Etude de l'ambiguïté des requêtes dans un moteur de recherche spécialisé dans l'actualité : exploitation d'indices contextuels

Fanny Lalleman

THÈSE

En vue de l'obtention du

DOCTORAT DE L'UNIVERSITÉ DE TOULOUSE

Délivré par : *l'Université Toulouse 2 Le Mirail (UT2 Le Mirail)*

Présentée et soutenue le *26/11/2013* **par :**

FANNY LALLEMAN

Étude de l'ambiguïté des requêtes dans un moteur de recherche spécialisé dans l'actualité : exploitation d'indices contextuels

---

JURY

PASCALE SÉBILLOT	PR, INSA Rennes	Président du Jury et Rapporteur
THOMAS LEBARBÉ	MCF HDR, Grenoble III	Rapporteur
LUDOVIC TANGUY	MCF HDR, Toulouse II	Examinateur
CÉCILE FABRE	PR, Toulouse II	Directeur
JOHANNES HEINECKE	Docteur, Orange	Invité

École doctorale et spécialité :
CLESCO : Sciences du langage
Unité de Recherche :
CLLE-ERSS (UMR 5263)
Directeur(s) de Thèse :
Cécile FABRE
Rapporteurs :
Pascale SÉBILLOT et *Thomas LEBARBÉ*

Étude de l'ambiguïté des requêtes dans un moteur de recherche spécialisé dans l'actualité : exploitation d'indices contextuels

Fanny Lalleman

Remerciements

Je tiens tout particulièrement à remercier Cécile Fabre, ma directrice de thèse, présente tout au long de cette thèse et qui m'a formidablement encadrée, aidée, et soutenue surtout lors des moments difficiles.

Je remercie aussi sincèrement, Johannes Heinecke, qui m'a accueillie et encadrée à Orange Labs. Il a su être présent et toujours disponible malgré la distance qui a pu nous séparer. Je remercie également Gilles Prigent de m'avoir permis de travailler dans de bonnes conditions.

Je suis également très reconnaissante à Pascale Sébillot et Thomas Lebarbé d'avoir accepté d'être les rapporteurs de ce travail de thèse, et à Ludovic Tanguy d'avoir bien voulu participer à mon jury de soutenance.

Je remercie l'ensemble du labo CLLE-ERSS pour ces belles années de thèse et plus particulièrement les membres l'axe TAL.

Je pense également à mes collègues de France Télécom (puis Orange) dont beaucoup ont rejoint d'autres horizons, Frédérique, Olivier, Gilles, Benoît, Michel, Edmond, Emilie, Aleksandra et bien d'autres. Je tenais à remercier Jean-Léon Bouraoui, pour sa présence, ses précieux conseils et ses relectures, tout comme Aurélie Picton et ses conseils survitaminés.

Je remercie également les doctorants de l'ERSS avec qui j'ai passé de très bons moments, Marie-France, Caro, François, Marine, Cécile, Stéphanie, Florian, Aurélie, Caitlin, Nikola, sans oublier les doc' Marianne, Clémentine, Lionel et bien sûr Christelle P. et j'en oublie bien d'autres.

Je remercie aussi les doctorants représentants avec qui j'ai vécu de nombreuses aventures. L'idex, que de souvenirs. Merci à l'école doctorale, de m'avoir permis de faire autant de choses, avec autant de cartes blanches. Clément, Agathe, Elsa, Fred, Caro, Marc, Marine, Etienne, Seb, Annelise et les autres, merci !

Il faut aussi que je remercie les amis qui ont subi ces années de thèse, mon manque de disponibilité. Je ne peux tous vous citer. Eric et Emma, Dany, Noémie, Thibault. Mathieu et Camille. Amélie aussi, toujours là quand ça ne va pas. Je voulais aussi remercier Sophie qui a animé nombre de mes soirées lannionaises ! Comment aurais-je fait sans ton accueil et ta bonne humeur ?

Impossible de remercier tout le monde avec ces quelques mots, mais je n'oublie pas le plus important. Ceux qui ont accepté de me voir moins souvent, d'être là quand il faut. Un merci « en vrai » ce sera bien mieux.

Table des matières

Table des figures

Liste des tableaux

Introduction

Dans le langage naturel, l'ambiguïté est une manifestation qui entraîne un flou interprétatif dû à la présence d'au moins deux sens concurrents et rend le lecteur ou l'interlocuteur indécis, à la recherche d'éléments complémentaires pouvant lever l'incertitude. Ce flou interprétatif est généralement dissipé par l'apport d'éléments désambiguïsateurs, orientant l'interprétation vers l'une des possibles significations.

Dans le contexte de la recherche d'information (RI), l'ambiguïté peut se manifester lorsqu'une requête est soumise à un moteur de recherche. Les requêtes sont des formulations artificielles, où les mots apparaissent dans un contexte linguistique très réduit. L'aspect artificiel des requêtes s'explique par le fait qu'elles sont construites dans un but précis afin d'interagir avec un environnement technologique. La taille des requêtes est en moyenne de 2 à 3 mots (Spink *et al.*, 2002b). Le contexte est donc absent. Par exemple, la requête *orange* est ambiguë. En effet, c'est un mot qui peut avoir plusieurs interprétations : un fruit, une couleur, une entreprise. Lorsque ce mot est utilisé pour chercher de l'information dans une base documentaire ou sur le Web, celui-ci peut ainsi devenir une requête ambiguë. La diversité des interprétations se manifeste alors parmi les résultats : le site Orange client, le site Orange business, la page Wikipédia consacrée à la couleur ou encore diverses informations sur le cours du fruit sur le marché mondial peuvent figurer dans les documents rapportés par un moteur généraliste.

Cette absence de contexte et la longueur très réduite des requêtes font qu'une requête ne peut pas être désambiguïsée dans un contexte langagier. Or, le problème de l'ambiguïté agit sur les performances du système de recherche (Schütze et Pedersen, 1995; Stokoe *et al.*, 2003; Spärck-Jones *et al.*, 2007; Sanderson, 2008). Car dans le cas d'une requête ambiguë, le système va avoir plus de difficultés à identifier le besoin informationnel sous-jacent à la requête. L'ambiguïté des requêtes est également un enjeu pour le CLIR (Cross-Language Information Retrieval) (Darwish et Oard, 2003). En effet, la traduction est très

dépendante de la désambiguïsation, d'autant plus quand le contexte est absent. L'ambiguïté des requêtes peut être aussi un facteur de la complexité des requêtes (Mothe et Tanguy, 2005). La complexité de ces requêtes peut amener certains à choisir d'exclure ces requêtes comme le fait par exemple Instagram [1]. Ce sont des requêtes comme *iphone*, *instagram* ou *popularpage*, considérées comme étant trop « génériques » et n'apportant pas d'information pertinente à l'utilisateur. La résolution de l'ambiguïté des requêtes est donc une plus value en terme de satisfaction utilisateur (Hearst, 2009).

Dans cette thèse, nous envisageons cette question de l'ambiguïté dans un domaine particulier qui est l'actualité. Notre thèse s'est déroulée dans le cadre d'une convention CIFRE avec Orange. Notre travail porte sur le moteur de recherche intégré au site 2424actu.fr, développé à Orange Labs. Le traitement de documents d'actualité pose des problèmes spécifiques dans la mesure où il s'agit d'un domaine particulièrement mouvant, comportant une proportion importante d'entités nommées. Notre but est de décrire les caractéristiques de la recherche d'information dans ce contexte afin de pouvoir mieux traiter la question de l'ambiguïté. Nous cherchons donc, à travers l'étude quantitative et qualitative d'un corpus de requêtes et de documents, à apporter des connaissances sur les spécificités du contexte de l'actualité et plus particulièrement sur les requêtes utilisateurs. Nous cherchons également à apporter des solutions au problème d'ambiguïté des requêtes et à tester de nouveaux dispositifs potentiellement déployables par Orange comme des modes de présentation des résultats différents.

En nous appuyant sur les travaux récents dans le domaine de la RI qui ont montré l'apport d'informations contextuelles pour mieux cerner et traiter plus adéquatement le besoin informationnel (Teevan *et al.*, 2005; Sanderson, 2008; Mothe, 2011) nous faisons l'hypothèse que les éléments d'information disponibles dans une application de RI (contextes présents dans la base documentaire, répétitions et reformulations de requêtes, dimension diachronique de la recherche) peuvent nous aider à examiner ce problème d'ambiguïté. Nous faisons également l'hypothèse que l'ambiguïté va se manifester dans les résultats ramenés par un moteur de recherche et que le signe de cette ambiguïté peut se matérialiser par une dispersion dans ces mêmes résultats.

Nos objectifs sont multiples. Nous voulons tout d'abord questionner la notion d'ambiguïté dans la pratique de la RI. De manière à comprendre la spécificité de ces manifestations dans ce contexte, nous les confrontons aux descriptions linguistiques de l'ambiguïté. Plus précisément, nous cherchons à étudier la

1. Pour lire le message du co-fondateur de Instagram à ce sujet sur un *groupe google* : http://ick.li/B9xzEn.

spécificité des requêtes d'un moteur d'actualité , dans un contexte caractérisé par une information mouvante et évolutive et des modalités de recherche spécifiques.

Nous souhaitons également pouvoir tester un dispositif de présentation des résultats révélant leur diversité. Ce dispositif doit être compréhensible par les utilisateurs et déployable dans un contexte industriel.

Pour répondre à ces objectifs, nous nous appuyons sur les indices contextuels présents dans l'environnement RI pour étudier les manifestations de l'ambiguïté. Ces éléments du contexte sont indispensables pour pouvoir reconstituer des élements interprétatifs autour de la requête. Ce sont évidemment des éléments qui ne remplacent pas un contexte langagier non artificiel, mais qui apportent des informations sur la manière dont s'est déroulé le parcours de recherche.

La thèse se découpe en trois parties. La première partie développe l'état de l'art sur l'ambiguïté des requêtes. La seconde partie décrit notre environnement d'expérimentation : le moteur de recherche spécialisé 2424actu et les données recueillies (requêtes et documents), la troisième et dernière partie s'intéresse à l'identification et l'analyse de l'ambiguïté des requêtes grâce à des indices contextuels sous l'angle de plusieurs expérimentations.

La première est consacrée à l'état de l'art sur l'ambiguïté en linguistique et en RI. L'absence d'un contexte langagier suffisant pour désambiguiser apparaît comme une question centrale et rend largement caduques les méthodes traditionnellement conçues pour traiter l'ambiguïté lexicale. Nous nous intéressons donc aux autres types de solutions développées ces dernières années. Nous introduisons une notion plus vaste du contexte, conçu cette fois non plus comme le co-texte d'apparition des mots de la requête mais comme l'ensemble des paramètres présents dans l'environnement de recherche, et susceptibles de venir préciser la nature du besoin informationnel. Le contexte qui entoure la tâche de recherche d'information peut être une source d'indices utile pour le repérage et la prise en charge de l'ambiguïté.

La deuxième partie est consacrée à la présentation de notre contexte d'expérimentation, à savoir le moteur de recherche 2424actu et ses paramètres contextuels disponibles dans l'application (chapitre 4). Dans le chapitre 5, nous présentons les données que nous avons constituées afin d'étudier l'ambiguïté des requêtes. Notre corpus comprend deux volets : le volet « document » et le volet « requêtes utilisateurs ». Ainsi après avoir discuté des contraintes applicatives liées à l'utilisation de données réelles et au contexte industriel, nous entamons une première caractérisation des requêtes à travers la mise en lumière des pra-

tiques de recherche. La caractérisation est une étape indispensable pour qualifier les données, permettant d'identifier leurs spécificités.

Dans la troisième partie de cette thèse, nous mettons en place plusieurs expérimentations cherchant à tester l'émergence de l'ambiguïté grâce à des indices contextuels. Dans le chapitre 6, nous testons l'apport d'une méthode de catégorisation des requêtes basée sur l'utilisation de catégories thématiques adaptées à l'actualité. Nous montrons que ces catégories, jusqu'ici exploitées uniquement pour organiser la base de documents s'avèrent plus adaptées qu'une source classique d'annotation de l'ambiguïté lexicale (Wikipédia). La pertinence d'une catégorisation experte, même rudimentaire, est ensuite évaluée du point de vue de l'utilisateur, dans le chapitre 7. Dans le dernier chapitre, plus exploratoire et basé sur une analyse de cas précis de requêtes populaires, nous faisons l'hypothèse que l'utilisation d'un faisceau d'indices contextuels (longévité des requêtes, proportion de versions étendues, diversité des contextes dans la base documentaire) fournit des informations sur le comportement d'une requête de manière globale.

Première partie

L’ambiguïté des requêtes : état de l’art

Chapitre 1

Caractérisation de l'ambiguïté en linguistique et en recherche d'information

L'ambiguïté des requêtes adressées à un moteur de recherche peut être envisagée comme un problème d'ambiguïté linguistique. En effet, les requêtes sont composées de quelques mots en langage naturel, souvent un ou deux mots isolés sans marque syntaxique. Toutefois, même si on parle de langage naturel dans les requêtes, ce sont des objets artificiels. Les requêtes sont des formulations façonnées et contraintes par l'environnement informatique auquel l'utilisateur se soumet. Le contexte linguistique est absent, il ne peut alors opérer sa fonction désambiguïsatrice. L'absence de contexte en RI réactive la discussion sur l'ambiguïté. Le préalable à cette recherche est une caractérisation de l'ambiguïté. Nous cherchons à savoir ce que peut apporter la linguistique sur la question de l'ambiguïté des requêtes.

Le but de ce chapitre est de savoir si l'ambiguïté telle que définie en linguistique correspond aux situations rencontrées en RI. Ce chapitre est donc structuré en deux temps. Nous commençons par caractériser l'ambiguïté en linguistique, en particulier les différents types d'ambiguïté qui ont été mis au jour par les linguistes et les tests qui permettent de les repérer. Nous parlons également du cas des noms propres qui sont touchés par l'ambiguïté et qui n'entrent pas forcément dans les types d'ambiguïté linguistique habituellement recensés. Dans le second temps de ce chapitre, nous nous focalisons sur l'ambiguïté des requêtes à proprement parler. Nous nous intéressons aux différents types d'ambiguïté décrits en RI. Nous voyons alors que la question de l'ambiguïté des requêtes est fortement liée aux méthodes utilisées pour les

repérer, méthodes qui seront exposées dans le chapitre suivant.

1.1 Caractérisation de l'ambiguïté

La caractérisation de l'ambiguïté doit nous permettre de mieux circonscrire le phénomène. En effet, quels types d'ambiguïté existent ? Comment ceux-ci se manifestent dans la langue ? Autant de questions qui nécessitent de caractériser l'ambiguïté et de la définir.

Fuchs (1994) définit qu'une expression est ambiguë « lorsqu'elle possède intrinsèquement plusieurs sens et qu'ils sont mutuellement exclusifs ». Par exemple, « défendre » est un terme ambigu, pouvant signifier « s'opposer formellement à la réalisation de » ou « protéger par la force » et cette deuxième signification peut être exprimée par d'autres formes comme *aider, protéger, secourir...* (Fuchs, 1996, p. 8). Selon Fuchs (1994), cette relation qui opère entre la forme et le sens d'une expression est indépendante des locuteurs et des situations car elle se situe au niveau du système de la langue. Une fois intégrée à un discours, le contexte rend la situation d'ambiguïté moins fréquente.

D'autre part, les ambiguïtés sont de formes différentes, par exemple, les ambiguïtés phonétiques (homophonie) comme pour *sein* et *saint*, ou les ambiguïtés morphologiques (*bois* : nom ou verbe conjugué) (Fuchs, 1994). Ces types d'ambiguïté sont au nombre de trois :

> « L'ambiguïté se manifeste sous trois formes, l'ambiguïté lexicale polysémique, l'ambiguïté lexicale homonymique, et l'ambiguïté structurale non lexicale » (Nicolas, 2006)

Dans la suite de ce chapitre, nous abordons la question de l'ambiguïté lexicale sous ses deux formes, la polysémie et l'homonymie. Puis dans un second temps, nous traitons le problème de l'ambiguïté structurale, qui se manifeste principalement au niveau de la syntaxe. Nous allons donc nous appuyer sur cette définition pour caractériser l'ambiguïté. Toutefois cette définition nous paraît ne pas prendre en compte tous les types d'ambiguïté, en particulier l'ambiguïté décrite par Croft et Cruse (2004), qui n'est ni lexicale, ni structurale, type d'ambiguïté qui mérite précision.

Par ailleurs en linguistique, on distingue les notions d'ambiguïté et de vague. Moeschler et Reboul (1994) vont jusqu'à les opposer. En effet, pour eux, « un terme est ambigu si l'on peut lui attribuer plusieurs extensions au moins partiellement différentes, alors qu'un terme est vague si l'on a des difficultés à déterminer précisément son extension. » (Moeschler et Reboul, 1994). L'impossibilité de décider si un terme est vague ou non est souvent liée au manque

de contexte. On peut également rapprocher la notion de vague de la notion d'indétermination. Dans un cas d'indétermination, « l'interprétation aboutit à la construction d'une signification globale univoque, mais excessivement pauvre » (Fuchs, 1989). Tout comme pour une expression vague, l'interprétation souffre du manque de contexte.

1.1.1 L'ambiguïté lexicale

Les deux formes d'ambiguïté lexicale, l'ambiguïté polysémique et l'ambiguïté homonymique, se distinguent principalement par l'origine de la source lexicale selon Croft et Cruse (2004). En effet, une unité homonymique dérive de deux sources lexicales distinctes, par exemple *avocat* peut être un fruit de l'arbre *avocatier* ou un homme de loi, dont le nom est emprunté au latin *advocatus* « homme dont la profession est de plaider » [1]. Une unité polysémique, a contrario, dérive de la même source lexicale. Ainsi par exemple, *patrimoine* peut désigner ce que des ascendants transmettent à leurs descendants (« patrimoine héréditaire », « transmettre son patrimoine »), ou bien désigner plus largement ce qui est considéré comme un *héritage commun* : « patrimoine archéologique », « patrimoine d'un peuple ». Ces deux significations proviennent toutefois de la même source lexicale et ont une étymologie commune.

1.1.1.1 L'ambiguïté homonymique

L'ambiguïté homonymique est aussi appelée ambiguïté « contrastive » par Pustejovsky (1995), ce qui souligne l'absence de lien entre les sources lexicales concernées pour l'homonymie. En ce sens, Riegel *et al.* (1998) déclarent que « deux mots sont déclarés homonymes si leurs paraphrases définitoires ne manifestent aucun trait sémantique commun ». Ainsi, on voit dans l'exemple suivant que les paraphrases définitoires de la forme *livre* n'ont aucun trait sémantique commun :

> *livre$_1$: Assemblage de feuilles en nombre plus ou moins élevé, portant des signes destinés à être lus.*
>
> *livre$_2$: Ancienne unité de poids, divisée en onces, variant selon les provinces de 380 à 552 grammes.*

Le critère diachronique est aussi mis en avant pour distinguer l'ambiguïté homonymique (Croft et Cruse, 2004). Cependant, ce critère n'est pas satisfaisant.

1. Les définitions et étymologies utilisées en exemple proviennent toutes du Trésor de la Langue Française http://www.cnrtl.fr/definition/

D'une part, il est difficile de déterminer l'étymologie de chaque forme ambiguë et, d'autre part, certaines formes sont homonymes tout en ayant une origine commune (comme l'exemple tiré de Riegel *et al.* (1998) *grève : arrêt collectif du travail* et *grève : plage de gravier*).

1.1.1.2 L'ambiguïté polysémique

La polysémie est un phénomène qui touche les unités du lexique. Selon Kleiber (1999), la polysémie se caractérise par une pluralité de sens liée à une forme, ces sens n'apparaissant pas totalement disjoints. Pustejovsky (1995) définit la polysémie comme étant une ambiguïté « complémentaire », où la catégorie morpho-syntaxique ne change pas et les multiples sens du mot se chevauchent et sont dépendants les uns des autres.

Essayer de définir la polysémie révèle certains problèmes comme celui de la définition du sens et sa délimitation. En effet, la question de la délimitation des sens varie selon les théories du sens et la finesse d'analyse qu'elles pratiquent (Riegel *et al.*, 1998). De ces variations découlent différentes visions de la polysémie : certains auteurs comme Kleiber (1999) définissent la polysémie en opposition à l'homonymie, tandis que d'autres comme Pustejovsky (1995) ou Victorri et Fuchs (1996) proposent l'existence d'un continuum entre ces deux types ambiguïtés lexicales. Cette conception de la polysémie place à un extrême les expressions monosémiques où le contexte ne joue aucun rôle. A l'autre extrême se situent les homonymes « purs ». Les auteurs considèrent alors que la polysémie est graduelle.

Enfin, la polysémie est un phénomène qui permet à la langue de créer et d'acquérir des nouveaux sens par le biais de figures rhétoriques comme la métaphore et la métonymie (Croft et Cruse, 2004). La métaphore consiste à utiliser un terme provenant d'un domaine concret pour exprimer une chose plus abstraite (Moriceau et Saint-Dizier, 2006), comme par exemple « virus informatique » qui est un emploi métaphorique du mot « virus » comme acteur biologique. La métonymie est également un processus de transfert de sens. Mais le sens métonymique entretient une relation de contiguïté avec le sens d'origine (Nunberg, 1995). Ainsi par métonymie, le *1er violon* désigne un violoniste (exemple de Jacquet *et al.* (2005)).

1.1.2 L'ambiguïté structurale

L'ambiguïté structurale est une ambiguïté manifestée par les constituants syntaxiques d'une phrase. Cette notion repose sur l'idée que chaque mot peut être

annoté par un label syntaxique. L'ambiguïté structurale peut être définie de la manière suivante :

> « Une forme manifeste une ambiguïté structurale si on peut lui faire correspondre au moins deux structures étiquetées distinctes. » (Nicolas, 2006)

Concrètement, ce type d'ambiguïté peut s'illustrer par l'exemple suivant de Riegel *et al.* (1998) (p. 126) : « La digestion du canard fut difficile ». Le complément *du canard* peut avoir deux rôles, actant sujet ou actant objet du verbe *digérer*. En effet, on peut comprendre que soit le canard a mal digéré (rôle actant sujet), soit quelqu'un a mal digéré le canard (rôle actant objet).

Riegel *et al.* (1998) soulignent également que l'ambiguïté structurale est souvent présente entre les fonctions de compléments du verbe et celles de compléments circonstantiels, comme on peut le voir dans cet exemple : « Les militaires rebelles se sont rendus en Argentine. » (*ibid*, p.141). *En Argentine* peut être un complément du verbe *se rendre* ou un complément de la phrase. Le test de détachement permet de valider la possibilité : « En Argentine, les militaires rebelles se sont rendus. ». L'ambiguïté structurale touche également les pronoms personnels et les adverbes comme *tout*, *encore*, *bien*, etc. qui peuvent être très ambigus dans le contexte de la phrase.

1.1.3 Les facettes sémantiques

Croft et Cruse (2004) proposent un troisième type d'ambiguïté lié à l'instanciation de différentes facettes sémantiques associées au mot. Les facettes décrivent le fonctionnement de certains mots qui ne relèvent pas de la polysémie traditionnelle et qui ont des sens ayant un degré d'autonomie important. Ainsi *Croft et Cruse (2004)* donnent en exemple les énoncés suivants où l'on peut observer que l'ambiguïté n'est pas attribuable au lexique ou à la syntaxe :

(a) *two books* : « book » désigne deux copies d'un même livre ou deux livres différents

(b) *two books in one* : « book » est un objet inclus dans un autre (deux livres en un)

(c) *a new book* : « book » peut désigner une nouvelle édition d'un vieux livre ou un livre écrit récemment

(d) *a long book* : peut être un livre qui contient beaucoup de mots ou qui a une forme non habituelle.

Les facettes n'ont pas le statut de sens lexical mais elles correspondent à des « types ontologiques » différents (*ibid*, p. 120). Par exemple, les facettes de *book*

sont [TEXT] (« a exciting book ; a difficult book ») et [TOME] (« a red book ; a dusty book »). Ces acceptions ne sont pas co-hyponymes, mais ce ne sont pas pour autant des sens à part entière. En effet, les facettes d'un seul mot ne sont pas toujours en compétition contrairement aux sens polysémiques ou homonymiques.

Croft et Cruse (2004) rapprochent ce fonctionnement sémantique des *qualia roles* proposés par Pustejovsky (1995). Ce sont quatre relations qui permettent de définir un objet ou une notion et ses relations avec le monde. Les *qualia roles* forment une structure d'interprétation à 4 niveaux : *constitutive role, formal role, telic role, agentive role*.

Cette rapide caractérisation des différents types d'ambiguïté linguistique montre que l'ambiguïté peut toucher différents niveaux linguistiques (le lexique ou la syntaxe).

1.1.4 Identifier l'ambiguïté

L'ambiguïté est diverse et difficile à appréhender. Un certain nombre de tests sont utilisés par les linguistes pour déterminer si une ambiguïté est présente ou non. Les tests cherchent à prouver que les sens portés par une même expression lexicale sont distincts les uns des autres, et qu'ils sont mutuellement exclusifs. La littérature propose deux types de tests que nous allons présenter ici. Les tests sont de deux sortes : les tests logiques qui s'appuient sur des jugements de valeur de vérité et les tests linguistiques qui s'appuient sur des jugements d'acceptabilité.

1.1.4.1 Les tests logiques

Guillon (1990, 2004) propose des « tests de jugement de valeur de vérité » pour juger de l'ambiguïté d'une phrase déclarative, c'est-à-dire qui déterminent si une phrase déclarative peut prendre deux valeurs alternatives. Ainsi pour Guillon (2004), il y a ambiguïté « si l'énoncé peut être à la fois vrai et faux ».

Nicolas (2006) illustre ce test par l'exemple suivant que nous reprenons : « Sylvain a vu un homme avec un télescope. ». Deux alternatives apparaissent pour interpréter cette phrase : soit Sylvain a observé un homme à l'aide d'un télescope, soit Sylvain a vu un homme transportant un télescope. Elle est donc ambiguë. Pourtant ce test ne permet pas de typer l'ambiguïté présente, il informe seulement sur la présence ou non d'un cas d'ambiguïté. En l'occurrence, c'est une ambiguïté structurale qui est présente dans cet exemple, marquée par la présence de deux structures syntaxiques différentes.

1.1.4.2 Les tests linguistiques

Les tests linguistiques proposés dans la littérature s'attachent principalement à tester des cas d'ambiguïté lexicale. Cruse (1986) distingue les tests « indirects » et « directs », distinction que nous allons conserver ici. Les tests indirects nécessitent de passer par une interprétation mettant en jeu des contextes différents, contrairement aux tests directs qui s'appliquent dans un contexte inchangé.

Les tests indirects pour détecter l'ambiguïté Cruse (1986) propose des tests utilisant les relations sémantiques qui lient les mots entre eux afin de prouver qu'une forme lexicale pointe vers différents sens. En effet, si une forme n'a pas les mêmes liens privilégiés en contexte, on peut supposer qu'elle prend des sens différents. Les tests proposés par Cruse (1986) reposent sur cette hypothèse. Nous considérons ici deux d'entre eux, le test de synonymie et d'antonymie.

Ces tests fonctionnent sur la même base. Chacun des tests consiste à remplacer la forme lexicale en question par son synonyme ou antonyme. Si celui-ci ne peut être le même dans tous les contextes où peut apparaître la forme lexicale testée, alors la forme porte plusieurs sens. Ces tests révèlent donc indirectement la présence d'une ambiguïté. En effet, si une forme ne peut être remplacée par une autre forme de sens similaire (synonymie) comme dans les exemples (1a) et (1b) testés en (1c) et (1d), ou antagoniste (antonymie) comme dans les exemples (2a) et (2b) testés en (2c) et (2d), alors la forme a plusieurs sens.

(1a) L'*avocat* parle à Jean.

(1b) Jean veut manger un *avocat*.

(1c) Le juriste ($\neq$ fruit*) parle à Jean.

(1d) Jean veut manger un fruit ($\neq$ juriste*).

(2a) La pièce a été peinte avec des couleurs *claires*.

(2b) Arthur a reçu des leçons de phonétique *claires*.

(2c) La pièce a été peinte avec des couleurs foncées ($\neq$ obscures*).

(2d) Arthur a reçu des leçons de phonétique obscures ($\neq$ foncées*).

Les tests directs pour détecter l'ambiguïté Les tests directs sont plus nombreux dans la littérature. Nous proposons ici de présenter les plus courants. Pour savoir si une forme est ambiguë entre deux sens possibles, on cherche à faire intervenir les deux sens à la fois.

Les tests reposent sur le fait que les différents sens d'une forme sont antagonistes et qu'ils ne peuvent apparaître simultanément dans le même contexte. Dans ce cadre, Cruse (1986) propose le test de coordination. Il permet de tester si les deux sens peuvent ou non être activés dans un contexte similaire, chose qui est très rare. L'exemple (3) est une application du test au verbe *expirer* qui peut prendre plusieurs sens. En l'occurrence, le test donne une phrase surprenante, où l'on associe la mort d'une personne à la fin de validité d'une autorisation.

(3) John et son autorisation *expirent* mardi prochain.

Un autre type de test possible est celui de l'anaphore (Nicolas, 2006; Cruse, 1986). Il consiste comme dans l'exemple (4) à reprendre le terme ambigu dans une seconde proposition par un pronom anaphorique. S'il n'est pas possible d'attribuer un sens_1 à *avocat* et un sens_2 au pronom *en* sans rendre l'énoncé étrange, alors la forme est ambiguë.

(4) Je ne veux pas d'*avocat*, Julie m'*en* a déjà vendu un.

Ces tests sont conçus pour être utilisés dans un contexte d'analyse sémantique. Ils reposent sur la possibilité de s'appuyer sur un contexte phrastique. Ils sont donc difficilement manipulables lorsque le contexte est réduit voire absent.

1.1.5 Le cas des noms propres

Le nom propre (NP) se distingue des noms communs par plusieurs aspects. Il peut en effet assurer deux types de fonction selon Biville (2005) : une fonction référentielle de désignation (prototypique) et une fonction de caractérisation lorsqu'il n'assume pas la fonction de désignation. La fonction de désignation permet d'identifier le référent visé, alors que la fonction de caractérisation est assumée par un ensemble de traits caractérisants (morphologique, syntaxique et sémantique). Ces traits se matérialisent par des constructions syntagmatiques comme *le perfide Ulysse* (exemple de Biville (2005, p. 39)) ou par des constructions de dérivés comme *un principe lacanien*.

Dans son emploi référentiel, comme le souligne Jonasson (1994), le NP désigne un individu en particulier. Cette propriété a engendré une vision du NP comme « désignateur rigide » (Kripke, 1980). Cependant, une deuxième vision du NP le qualifie au contraire de « désignateur souple » (Paveau, 2008), porteur d'une « signifiance » (Siblot, 1987). Cette vision envisage le NP en discours. Cette opposition théorique doit toutefois être complétée par l'étude des NP, qui présentent une grande diversité. En effet, il existe plusieurs types de NP. Jonasson (1994) en distingue deux :

– les noms propres « purs » : ce sont des noms propres de personnes ou de lieux, « des particuliers auxquels on attribue toujours un NP et dont on suppose qu'ils ont un NP ». Les NP purs seraient des NP comme *Paul*, *Paris* ou *Hollande*.
– les noms propres « descriptifs » : ils sont distingués du point de vue morphologique par Jonasson (1994). Ce type de NP est composé majoritairement de noms communs, parfois associés à des modificateurs adjectivaux ou prépositionnels, comme par exemple *Académie Française* ou l'*Assemblée Nationale*.

Selon Jonasson (1994), les NP purs ont tendance à avoir des emplois non rigides, contrairement aux NP descriptifs qui constituent une « véritable description du particulier qu'ils désignent ». C'est le rôle de désignateur rigide qu'est censé assumer le NP. Or, cette fonction n'est pas réellement assurée par les NP de lieux et de personnes selon Jonasson (1994), la mono-référentialité étant plus caractéristique des NP descriptifs que des NP purs. En considérant la question de l'ambiguïté, c'est donc la caractéristique portée par les NP purs (NP de lieux et de personnes) qui va particulièrement nous intéresser.

La polyréférentialité du nom propre La notion de « polyréférentialité » (Gary-Prieur, 2001) s'inscrit dans cette vision du NP comme désignateur souple. Elle rend compte de l'existence d'une pluralité de référents associés à l'entité désignée par le NP, ces référents gardant des liens entre eux. A la suite de Gary-Prieur (2001), Lecolle (2004) affine la notion de polyréférentialité et distingue au final deux niveaux de la polyréférentialité : interne et externe.

La polyréférentialité interne désigne le caractère composite de l'ensemble référent (référent discursif), ce qui correspond à un « épaississement » du NP (Lecolle, 2004). Les éléments sont inextricablement liés, comme dans l'exemple (6) tiré de Lecolle (2004). *Cannes* prend une valeur évènementielle qui se combine aux acteurs de l'évènement. Il n'est alors pas possible de distinguer ces deux valeurs.

(6) Comment *Cannes* constitue son affiche, en l'espace de neuf mois. (En titre, Libération 09/05/2001)

La polyréférentialité externe se manifeste par la pluralité de référents possibles, pouvant aboutir à une concurrence d'interprétations. Ce type de polyréférentialité met en évidence le caractère flexible et dépendant du contexte, de la référence portée par le NP (ou de la fonction référentielle du NP). Cette situation est illustrée dans l'exemple (7), emprunté à (Lecolle, 2004) où le NP *Washington* est touché par une polyréférentialité externe. Il peut prendre en

effet deux valeurs (en plus de sa valeur locative) : celle d'actant institutionnel et celle d'évènement [2].

(7) 200 000 manifestants anti-mondialisation veulent paralyser *Washington*. (Le Monde, 28/09/2002).

Cette capacité observée des NP à d'une part « s'épaissir » du point de vue référentiel et d'autre part à se fragmenter en une pluralité de domaines référentiels nous incite à retenir les NP comme pouvant être porteurs d'ambiguïté.

1.1.6 Conclusion

L'ambiguïté en linguistique est donc un phénomène rare grâce à l'intervention du contexte linguistique. Les ambiguïtés causées par le manque de contexte linguistique ne sont pas désignées comme de l'ambiguïté mais comme une « indétermination ». En effet, comme le précise Fuchs (1994) les différentes significations misent en jeu sont exclusives, obligeant le locuteur à choisir entre les significations concurrentes.

Nous avons également vu que les tests proposés pour identifier l'ambiguïté s'appuient sur le contexte linguistique pour opérer à la manière d'un « filtre ». Ces tests forcent le rôle que le contexte est censé jouer dans des cas où il n'est pas performant. Ils permettent également d'identifier les formes qui peuvent être porteuses de significations différentes.

Enfin, le cas particulier du nom propre s'avère particulièrement important. Alors qu'il est considéré comme un désignateur rigide, la réalité s'avère bien plus complexe. Les NP peuvent être polyréférentiels et être sémantiquement flexibles. Ces conclusions nous amènent à nous demander comment cette catégorisation va s'intégrer au contexte de la recherche d'information.

1.2 L'ambiguïté des requêtes en Recherche d'Information

Rappelons que la requête est un usage totalement artificiel de la langue. Elle se présente sous la forme de mots juxtaposés et ne dépasse pas en général 2 à 3 mots (Spink *et al.*, 2002b). Ces formulations réduites reposent la question de l'ambiguïté. La fonction naturelle de « filtre », qui est censée être assurée par le contexte linguistique est absente. La situation de recherche d'information est donc propice à l'apparition d'ambiguïtés. Par conséquent, on peut s'attendre à rencontrer en RI deux types d'ambiguïtés. Un premier type est inhérent à

2. L'évènement renvoie à une réunion des ministres des finances du G7 du 27 septembre 2002 selon (Lecolle, 2004)

l'ambiguïté présente dans la langue comme le définit Fuchs (1994) et non filtrée par le contexte linguistique, elle peut produire de l'ambiguïté lexicale. Un deuxième type d'ambiguïté peut être induit par le manque de contexte linguistique.

1.2.1 L'ambiguïté lexicale des requêtes

De nombreux travaux assimilent l'ambiguïté présente dans les requêtes des utilisateurs à de l'ambiguïté lexicale. Le recours aux ressources lexicographiques permet une détection de l'ambiguïté lexicale à grande échelle. Les différents sens présents dans chacun des mots d'une requête sont repérés grâce au concours de ressources lexicales telles que les dictionnaires (Krovetz et Croft, 1992) ou les encyclopédies (Sanderson, 2008; Santamaría *et al.*, 2010; Welch *et al.*, 2011). Cette démarche s'inscrit dans le domaine de la désambiguïsation automatique que nous développerons dans le chapitre 2.

1.2.1.1 L'ambiguïté lexicale à la lumière des ressources lexicographiques

Les premiers travaux sur l'ambiguïté des requêtes s'inscrivaient dans le domaine de la désambiguïsation automatique. Les premiers travaux s'appuyaient donc sur des dictionnaires pour repérer l'ambiguïté comme Krovetz et Croft (1992) avec le Longman Dictionary of Contemporary English.

WordNet a été ensuite utilisé par des travaux qui cherchaient à désambiguïser des requêtes (Voorhees, 1993; Gonzalo *et al.*, 1998). Or, ce type de méthode n'est pas capable de déterminer le sens porté par les mots de la requête (Sanderson, 2000). WordNet présente un deuxième inconvénient pour cette tâche, les noms propres ne sont pas présents dans la ressource.

L'inadéquation de WordNet a largement contribué à l'utilisation d'une ressource plus importante, l'encyclopédie Wikipédia. En effet, Wikipédia propose un étiquetage des mots ambigus, qui sont par défaut des cas d'homonymie. Les travaux qui se basent sur l'encyclopédie en ligne utilisent les pages dites de « désambiguïsation ». Ce sont des listes qui font référence à des mots, des expressions ou des noms propres ambigus. Par exemple, l'article *Chicago* est une grande ville des USA mais le mot *Chicago* peut référer à d'autres choses : la rivière Chicago, la comédie musicale *Chicago* ou encore le film du même nom. L'article *Chicago* débute donc par la chaine {{Voir homonymes|Chicago (homonymie)}}. Et un lien renvoie vers le deuxième type de page qui liste les différents articles portant le même nom, soit la page Chicago_homonymie.

Sanderson (2008), tout comme Song *et al.* (2009) utilisent cette méthode. Leurs résultats sont comparables sur un corpus de requêtes issues du moteur Live Search, avec 16% de requêtes considérées comme ambiguës. Pour obtenir ces résultats, ils ont également utilisé en complément WordNet (Fellbaum, 1998) pour Sanderson (2008) et le dictionnaire TheFreeDictionary pour Song *et al.* (2009). Sanderson (2008) a également analysé les requêtes du moteur UK's Press Association, et 23,6% de celles-ci s'avèrent être ambiguës à la lumière des ressources mobilisées.

Ces résultats nous informent sur deux aspects. D'une part, ils confirment que les requêtes comportent des mots ambigus du point de vue de ces ressources lexicographiques et encyclopédiques. D'autre part, les résultats obtenus à partir des requêtes des deux moteurs suggèrent que l'ambiguïté varie selon les requêtes étudiées et leur provenance, les requêtes du moteur UK' Press étant plus nombreuses à avoir été diagnostiquées comme ambigues.

1.2.1.2 L'ambiguïté lexicale à la lumière de la base documentaire

Pourtant plusieurs questions se posent. Comme Spärck-Jones *et al.* (2007) le précisent, si les requêtes très courtes, en particulier celles ne comportant qu'un seul mot, sont ambiguës, elles ne correspondent pas pour autant aux sens présents dans les dictionnaires. Et ce, pour au moins deux raisons.

D'une part, les requêtes contiennent un grand nombre de noms propres, à tel point que c'est devenu l'argument principal pour utiliser Wikipédia comme ressource. L'étude des requêtes utilisateurs de AllTheWeb et Altavista a mis en évidence que 11 et 17% des requêtes étaient composées d'un nom propre désignant une personne (Spink *et al.*, 2004). Gan *et al.* (2008) montrent dans une étude des logs de requêtes de AOL (2006) qu'il y a environ 38% des requêtes qui contiennent des noms désignant des lieux comme *New York*. Or, les noms propres peuvent être une source de polyréférentialité comme nous venons de le voir dans la section 1.1.5.

D'autre part, l'ambiguïté des requêtes ne doit pas être considérée seulement à la lumière de ressources lexicographiques. Ce sont les documents auxquels les requêtes permettent d'accéder qui éclairent sur le sens des requêtes. Les travaux de Rahurkar *et al.* (2008) sur des requêtes [3] du moteur MSN le montrent. En effet, même s'ils ont déterminé que 85% des sens portés par les requêtes de leur corpus étaient couverts par Wikipédia, la couverture n'est pas la même pour toutes les requêtes. Par exemple, elle est bien plus basse pour une requête

3. Ce sont des requêtes mono-mots déjà utilisées par Zeng *et al.* (2004)

comme *jordan*, et ce pour une raison liée à un défaut de correspondance entre les sens présents dans la base documentaire et ceux recensés dans Wikipédia. En effet, pour cette requête Wikipédia ne recense pas tous les Michael Jordan, professeur ou autres entreprises du même nom, que l'on retrouve dans les résultats à la requête.

On peut donc poser la question de l'inadéquation des ressources, à la fois à propos de la question de l'ambiguïté et à propos de la prise compte du contexte de RI.

1.2.2 Vers un autre type d'ambiguïté des requêtes : les requêtes larges

Rompant avec la vision lexicographique de l'ambiguïté des requêtes, certains travaux proposent d'analyser d'autres formes d'ambiguïté. Ils avancent que des requêtes peuvent être ambiguës si elles renvoient à plusieurs « sous-sujets » (Zhai *et al.*, 2003) ou à plusieurs « facettes » (Hearst, 2006). Spärck-Jones *et al.* (2007) précisent que l'ambiguïté touchant les requêtes peut être de nature lexicale ou référentielle :

> « The ambiguity may be of word *sense*, or of reference *aspect*. The request « house » may mean 'building', 'home' or 'firm' and the request « house prices » may refer to actual prices or economic factors. There is also the issue of request *type* e.g. topic vs home-page seeking. »

L'ambiguïté des requêtes serait double, à la fois une ambiguïté linguistique touchant le sens des mots d'une requête, et à la fois une diversité référentielle qui se manifesterait par la capacité d'une requête à être composée de différents aspects. Clarke *et al.* (2009) l'illustrent de la manière suivante :

- une ambiguïté de nature lexicale : la requête est « ouverte » à différentes interprétations. Par exemple, la requête *java* en anglais peut signifier *programming language*, *coffee*, ou encore *island* etc.
- une ambiguïté de nature référentielle : l'utilisateur peut s'intéresser à différents « aspects » contenus dans une requête pour un même sens. Dans ce cas, la même requête *java* va renvoyer à des *development kit download*, *courses*, *books*, *language specifications*, ou encore des *tutorials*.

On retrouve ce type de distinction vis-à-vis de l'ambiguïté des requêtes dans Song *et al.* (2009). Ils utilisent une typologie à trois éléments qu'ils ont définie dans le but d'annoter manuellement des requêtes :

- les requêtes ambiguës (type A) : ce sont des requêtes qui ont plus d'un sens, par exemple la requête *giant* a plusieurs référents qui sont *Giant Company*, *Giant* (film), *San Francisco Giant* (équipe de basket ball) ;

– les requêtes « larges » (type B) : ce sont des requêtes qui couvrent plusieurs sujets ou thématiques. La requête *songs* va alors ouvrir vers *song lyrics*, *love songs* ou encore *download songs ;*
– les requêtes non ambiguës (type C) : ces requêtes ont un sens spécifique et un référent facilement identifiable comme par exemple *Billie Holiday* (chanteuse jazz).

Le type A décrivant des requêtes ambiguës est à rapprocher du cas d'ambiguïté lexicale décrite par Clarke *et al.* (2009) et en 1.2.1. De même, le type B désignant des requêtes larges est très proche du deuxième cas décrit par Clarke *et al.* (2009). La requête peut recouvrir plusieurs aspects et ouvrir vers une diversité de sujets. On retrouve également cette distinction entre requêtes ambiguës et « larges » chez Zeng *et al.* (2004) qui considèrent que les requêtes de type « générique » recouvrent de nombreux « sous-sujets » comme *maps*, *chat* ou *flower* (en anglais).

Toutefois, les critères utilisés par Song *et al.* (2009) pour discriminer les requêtes de type ambiguë (type A) et non ambiguë (type C) ne sont pas clairement établis. En effet, potentiellement l'exemple donné comme requête non ambiguë (*Billie Holiday*) peut être discuté, puisqu'il s'avère qu'il existe un album éponyme de Billie Holiday. Si la qualification d'une requête comme ambiguë ou non peut être discutable, cela montre surtout que le problème de l'ambiguïté d'une requête doit être envisagé dans le cadre de la base documentaire à laquelle elle donne accès. En l'occurrence, la requête *Billie Holiday* sera ambiguë si dans la base documentaire sont mentionnés à la fois la chanteuse et son album éponyme.

La distinction entre « ambigu » et « non ambigu » est donc difficile à établir, laissant apparaître des requêtes au statut intermédiaire. Song *et al.* (2009) les distinguent en créant une catégorie de requêtes contenant un terme générique (requête de type B). Ils décrivent un phénomène d'ambiguïté proche de la sous-spécification, sens vague, indétermination ou généralité. Cette forme d'ambiguïté décrit un sens général ou inclusif qui peut avoir différentes significations spécifiques selon le contexte, par exemple *vache* va pouvoir prendre un sens général « animal ruminant » ou un sens spécifique « femelle du taureau ». D'autres comme Chirita *et al.* (2005), les définissent comme étant « semi-ambiguës ». Selon ces auteurs, une requête semi-ambiguë aurait deux ou trois sens alors qu'une requête ambiguë aurait au minimum trois sens. Toutefois cette définition est proposée au regard d'une ressource conçue pour structurer le web et des documents retournés par un moteur de recherche (Google Directory) et n'est donc pas réutilisable dans un autre contexte de RI.

1.2.3 Caractériser l'ambiguïté des requêtes : proposition d'une synthèse

Nous listons plusieurs problèmes qui freinent la caractérisation de l'ambiguïté des requêtes. L'ambiguïté des requêtes est plus difficile à établir que l'ambiguïté dans le discours. La vision lexicographique du problème d'ambiguïté des requêtes masque la réalité de la situation. Les ressources projettent une ambiguïté *a priori* qui est artificielle et ne reflètent pas la réalité des emplois des termes de la requête dans la base documentaire.

L'ambiguïté des requêtes ne se limite pas à l'ambiguïté lexicale, comme on a pu le voir avec le cas des requêtes « larges ». Celui-ci met en exergue le fait qu'une requête peut comporter plusieurs aspects et être porteuse de diversité. Ce type d'ambiguïté semble être causé par le manque de contexte qui crée une situation d'indétermination.

De plus, la caractérisation ne peut être envisagée sans considérer le problème de l'ambiguïté au regard de la base documentaire à laquelle elle donne accès. Une autre manière de définir l'ambiguïté des requêtes pourrait donc venir des travaux qui envisagent l'ambiguïté en fonction du comportement des utilisateurs (Dou *et al.*, 2007; Teevan *et al.*, 2008; Wang et Agichtein, 2010). Une diversité de comportements pour une même requête serait alors le signe d'une ambiguïté comme nous le verrons en 2.2.2.4 et en 3.2.3.

1.3 Conclusion

L'éclairage apporté par la linguistique à propos de la question de l'ambiguïté des requêtes est important. Il nous permet en effet de mieux comprendre comment les ambiguïtés se forment en RI. La linguistique précise également le rôle joué par le contexte et les conséquences de son absence. Cependant, l'environnement RI crée une situation différente. Toute ambiguïté portée par un mot d'une requête se retrouve confrontée, non pas, au filtre contextuel d'un discours mais à celui d'une base documentaire comportant un grand nombre de contextes linguistiques variés.

De plus, l'absence de contexte en RI crée une situation génératrice d'ambiguïté ou d'indétermination. Ce type d'ambiguïté ne peut être recensé de façon pérenne, ni être prévisible par avance. La vision lexicographique du problème de l'ambiguïté pose plusieurs problèmes. En effet, les ressources lexicographiques ne sont peut être pas les mieux indiquées pour comprendre l'ambiguïté (Véronis, 2001, 2002), mais c'est aussi une démarche qui se prive du filtre

contextuel que peut apporter la base documentaire. Ainsi, les travaux qui ont considéré le problème de l'ambiguïté des requêtes du point de vue des résultats de recherche ont pu avoir une vision plus globale sur cette question.

Le problème clé de l'ambiguïté des requêtes est donc la question du contexte linguistique manquant. Pour répondre à cette question, nous allons examiner les solutions trouvées à ce problème dans le chapitre 2. Puis dans le chapitre 3, nous nous intéressons aux éléments contextuels présents et utilisables en RI.

Chapitre 2

Traitement de l'ambiguïté en recherche d'information

Nous venons de voir dans le chapitre 1 que même si l'ambiguïté des requêtes ne se limite pas à de l'ambiguïté lexicale, une grande partie des travaux sur l'ambiguïté des requêtes la traite comme telle. Cette vision de l'ambiguïté amène du point du vue automatique à utiliser des méthodes de désambiguïsation. C'est pourquoi nous proposons dans un premier temps de revenir sur la désambiguïsation lexicale automatique, et les processus des différentes méthodes de désambiguïsation. La désambiguisation lexicale est une tâche qui consiste à identifier le sens des mots en contexte et de manière automatique. Elle est considérée comme un « problème d'intelligence artificielle complet » (Ide et Véronis, 1998; Navigli, 2009). Nous proposons un paranoma des approches existantes, ce qui nous permettra ensuite d'aborder la question de l'ambiguïté en Recherche d'Information. Les mauvais résultats de ces méthodes ont ouvert de nouvelles pistes de recherche que nous explorons dans un second temps. Les solutions proposées et expérimentées sont de deux sortes : les solutions qui s'appliquent directement sur les requêtes et les solutions qui s'appliquent aux résultats obtenus à partir des requêtes ambiguës. Nous allons détailler ces deux types d'approches qui ont en commun de ne pouvoir être mises en oeuvre sans la prise en compte d'informations contextuelles.

2.1 La désambiguïsation lexicale : définition et méthodes

La désambiguïsation lexicale (*Word Sense Disambiguisation*) intervient dans la plupart des applications de traitement automatique de la langue : traduction

automatique, recherche d'information, analyse syntaxique, synthèse vocale, etc. (Ide et Véronis, 1998).

La désambiguïsation lexicale permet selon Ide et Véronis (1998) d'associer un mot donné dans un texte ou un discours avec son sens, tout en le distinguant des autres sens existants pour ce mot. D'après Kilgarriff (2006) cela revient à assigner au mot un sens « prédéfini » dans un dictionnaire ou un lexique. Toutefois, c'est une tâche difficile, les phénomènes de polysémie et d'homonymie étant difficiles à délimiter manuellement comme on l'a vu dans le chapitre 1. Donc, a fortiori, la difficulté persiste dans la démarche d'automatisation.

La désambiguïsation lexicale est réalisée majoritairement grâce à deux types de méthodes : les approches basées sur des ressources lexicographiques et les approches basées sur des corpus. L'évolution des recherches dans le domaine de la désambiguïsation a vu l'émergence d'approches dites « mixtes » qui empruntent aux méthodes à base de ressources lexicales et à celles basées sur corpus. Après avoir discuté ces deux types de méthodes, nous allons nous focaliser sur l'évaluation de la désambiguïsation automatique, ce qui nous permettra de mieux comprendre la difficulté d'une telle tâche.

2.1.1 Les méthodes pour désambiguïser

Nous proposons ici de présenter les deux types de méthodes utilisées pour la désambiguïsation lexicale automatique, en distinguant d'une part les méthodes faisant usage de ressources lexicographiques et d'autre part les méthodes basées sur corpus. Nous allons voir que les méthodes ne se distinguent pas seulement du point de vue des ressources utilisées, mais également du point de vue de la démarche. En effet, si les premières cherchent à retrouver des sens déjà catalogués et typés dans une ressource, les secondes cherchent à faire émerger les sens présents dans les corpus étudiés. Enfin, nous présentons les travaux qui font appel aux deux types de méthodes, et qui combinent ressources lexicales et corpus.

2.1.1.1 Approches basées sur des ressources lexicographiques

Les approches qui se basent sur des ressources lexicographiques sont nombreuses (pour une revue exhaustive des travaux voir Ide et Véronis (1998); Navigli (2009)). Les travaux en désambiguïsation ont en effet exploré le potentiel d'un grand nombre de ressources lexicales de dictionnaires, thésaurus, réseaux sémantiques. Nous allons présenter leurs différents points forts et difficultés.

Les dictionnaires Les dictionnaires informatisés (*Machine-Readable Dictionaries*) ont été la principale ressource utilisée dans les années 1980. Ils continuent de fait à être une source d'information essentielle en ce qui concerne le sens des mots. Les travaux développés à partir des dictionnaires vont s'appuyer sur un principe simple : « lorsque plusieurs mots sont cooccurrents, le sens le plus probable pour chacun de ces mots est celui qui maximise ses relations avec le sens des mots cooccurrents » (Audibert, 2003).

Lesk (1986) a proposé le principal algorithme basé sur l'utilisation d'un dictionnaire. Pour deux mots donnés (w_1, w_2), les sens des mots cibles sont les définitions qui ont le plus haut taux de recoupement. Cet algorithme s'appuie sur la création d'une base de connaissances qui comporte pour chaque sens d'un mot la liste des mots apparaissant dans sa définition. Une telle méthode est intéressante par sa capacité à détecter finement les sens. Toutefois, elle demande la création de ressources importantes, qui n'offrent souvent qu'une couverture lexicale faible.

Les thésaurus Les thésaurus permettent d'avoir des informations sur les relations sémantiques existant entre les mots comme la synonymie, l'antonymie et des relations de hiérarchie (Kilgarriff et Yallop, 2000). Ils se caractérisent par une structure hiérarchique en conformité avec une norme, les concepts étant reliés par des relations sémantiques normalisées (Lefèvre, 2000).

Parmis les travaux de désambiguïsation utilisant des thésaurus, on peut citer par exemple les travaux de Yarowsky (1992) qui se basent sur l'utilisation du Roget's Thesaurus (Roget, 1911). Ce type de méthode donne une précision élevée (92%) mais ne désambiguïse efficacement que les noms (Yarowsky, 1992). Selon Audibert (2003), les thésaurus combinent donc les mêmes inconvénients que les dictionnaires. Ils manquent de couverture et de cohérence.

Les réseaux sémantiques Dans les années 1980, de grandes bases de connaissance ont été constituées manuellement. Le réseau sémantique le plus connu est WordNet (Fellbaum, 1998). Ce réseau a toujours suscité un grand intérêt pour la communauté du Traitement Automatique des Langues, du fait de sa disponibilité et de son unicité. Pour rappel, WordNet est organisé autour de *synsets*, « synonymes » auxquels sont rattachés les mots, ordonnant les noms, adjectifs et adverbes de la langue anglaise en ensemble de synonymes. Pour compléter la relation de synonymie, différentes relations lexico-sémantiques sont utilisées pour relier les mots entre eux comme la relation de hiérarchie (hypéronymie et hyponymie) et la relation de tout-partie (holonymie et meronymie).

L'arrivée de WordNet a permis l'introduction de nombreuses mesures de similarité sémantique, exploitant le réseau proposé. Les méthodes les plus simples ont cherché à calculer la distance qui existe entre deux sens dans WordNet (Rada *et al.*, 1989). Selon Navigli (2009), une densité élevée de concepts, unis par des liens d'hyponymie, serait l'indication d'un sens plus « probable » en contexte, permettant d'orienter le choix à faire pour désambiguïser.

Toutefois, l'utilisation même de WordNet reste complexe et comme le signale Navigli (2009), beaucoup de travaux sous-exploitent la ressource, n'utilisant que les relations hiérarchiques présente dans WordNet.

2.1.1.2 Approches basées sur corpus

On distingue deux types d'approches se basant sur des corpus, qui peuvent être étiquetés ou non. Les premières font intervenir des corpus étiquetés alors que les secondes adoptent une démarche inductive. Les méthodes d'apprentissage supervisé font intervenir des corpus enrichis par diverses annotations, et comprennent des phases d'apprentissage (par règles ou par exemplification). Les approches sur corpus non étiquetés font appel à des méthodes d'apprentissage non supervisé, qui ne sollicitent aucune ressource extérieure.

Approches sur corpus étiquetés Les travaux basés sur corpus étiquetés s'appuient sur deux types d'étiquetage : morpho-syntaxiques et sémantiques. Les étiquetages syntaxiques ont été utilisés par certains travaux, ceux-ci ont basé leur phase d'apprentissage sur des dépendances syntaxiques telles que sujet-verbe, verbe-objet, adjectif-nom, etc (Lin, 1997; Pantel et Lin, 2002). Les dépendances syntaxiques permettent d'avoir une vue « enrichie » du contexte. Nous présentons ici deux types de corpus étiquetés qui ont été utilisés pour des travaux de désambiguïsation : les corpus alignés et les corpus étiquetés sémantiquement.

Les progrès de l'étiquetage syntaxique (Gale *et al.*, 1993) ont encouragé l'apparition des corpus alignés bilingues. Ces corpus permettent d'avoir un mot et sa traduction exacte dans une autre langue ce qui désambiguïse le mot en question. Par exemple, le mot *pen* en anglais peut se traduire selon le contexte en *stylo* ou *enclos* en français (Ide et Véronis, 1998). Cette méthode simple se heurte toutefois aux problèmes d'ambiguïté qui persistent dans les langues servant à la désambiguïsation comme le signalent Ide et Véronis (1998). Le deuxième problème de cette méthode est la faible disponibilité de ce type de corpus, hormis les corpus spécialisés comme les débats bilingues du parlement canadien.

La création du premier corpus annoté sémantiquement a permis la réalisation de travaux à base de méthodes statistiques. Le corpus SemCor est créé en 1993 (Miller *et al.*, 1993), à partir du Brown Corpus (Francis et Kucera, 1982). 234 000 occurrences ont été annotées manuellement : catégories grammaticales, lemmes et sens provenant de WordNet 1.6. Il existe également le DSO corpus (Ng et Lee, 1996) qui contient environ 192 800 occurrences : 121 noms et 70 verbes, les plus fréquents et les plus ambigus de l'anglais. Les occurrences ont été annotées avec WordNet 1.5. Ces corpus annotés sémantiquement restent limités mais ont permis de nombreuses évaluations dans le cadre des campagnes d'évaluation Senseval puis Semeval (ce que nous développerons en 2.1.2).

Les algorithmes de désambiguïsation à partir de corpus étiquetés fonctionnent en deux temps. Tout d'abord, la phase d'apprentissage permet d'extraire les connaissances nécessaires, en l'occurrence, un sens par lexème. Et une fois l'apprentissage réalisé, dans un deuxième temps, les algorithmes sont en capacité d'assigner un sens à chaque lexème rencontré (Audibert, 2003).

Ces méthodes s'avèrent performantes dans les campagnes d'évaluation telles que Senseval (Navigli, 2009). Cependant, leur utilisation reste dépendante des ressources étiquetées, en particulier au niveau sémantique où l'étiquetage est encore manuel.

Approches sur corpus non étiquetés Face à des ressources limitées et des corpus étiquetés assez restreints, les méthodes d'apprentissage non supervisées permettant de travailler sur des corpus non étiquetés sont apparues comme une alternative possible. Ces approches sont basées sur l'idée que le sens particulier d'un mot apparaît dans des contextes similaires. En classifiant automatiquement les occurrences d'un mot dans un texte, procédé appelé *clustering*, des « groupes » de contexte similaires se forment (Navigli, 2009). Ainsi, ces approches déduisent qu'un sens particulier est supposé correspondre à un « groupe » défini automatiquement. Ces approches non supervisées réussissent à discriminer les sens des mots en contexte, mais elles ne découvrent pas des clusters équivalents aux sens traditionnellement inventoriés dans les dictionnaires .

Les deux principales approches à base de clustering se basent soit sur le calcul de la distance entre les mots (Pedersen et Bruce, 1997; Schütze, 1992), soit sur le calcul de la connexion entre les mots (Véronis, 2003). Schütze (1992) est un des premiers à avoir utilisé une méthode de clustering basée sur une représentation vectorielle des mots et donc sur la distance qui sépare deux mots. Ce type de méthode repose sur l'hypothèse suivante :

> « The underlying hypothesis of this model is that the distributional profile of words implicitly expresses their semantics. » (Navigli, 2009)

La seconde approche consiste à représenter les paires de mots en connexion lorsqu'elles cooccurrent dans une relation syntaxique, à un niveau local (dans le même paragraphe), ou dans un large contexte. Ce type de méthode construit un graphe basé sur les relations syntaxiques comme le fait l'algorithme HyperLex de Véronis (2003). Pantel et Lin (2002) ont appliqué ce type de méthode avec l'algorithme CBC (*Clustering By Committee*) dans une tâche d'identification des sens issus de WordNet. L'algorithme réussit à atteindre 61% de précision et 51% de rappel, toutefois l'évaluation ne semble pas adaptée puisque, comme le souligne Navigli (2009), les sens découverts par le clustering ne sont pas équivalent à ceux répertoriés dans les dictionnaires.

L'évaluation de telles méthodes pose donc des problèmes, puisque les étalons habituels ne sont plus utilisables. Par ailleurs, les différentes méthodes de classification ont besoin d'un grand volume de données pour obtenir des résultats intéressants (Navigli, 2009).

2.1.1.3 Combiner les connaissances structurées et les corpus

Le *bootstrapping* désigne une méthode qui répond à la fois au problème des approches sur corpus étiquetés et non étiquetés. En effet, le bootstrapping consiste à entrainer un classifieur sur un corpus annoté, puis à étendre ce corpus grâce à l'apprentissage initial. Ainsi, à chaque nouvelle itération, l'algorithme engrange de nouveaux exemples étiquetés. Les travaux utilisant le bootstrapping en désambiguïsation peuvent utiliser un ou plusieurs classifieurs. Par exemple Yarowsky (1995) utilisait un seul classifieur qu'il autoentrainait en se basant sur deux heuristiques : « un sens par collocation » (Yarowsky, 1993) et « un sens par discours » (Gale *et al.*, 1992). La première heuristique repose sur l'hypothèse que les proches voisins d'un mot contribuent à déterminer le sens de ce mot. La seconde indique qu'un mot réfère au même sens dans un discours ou document donné.

Les premiers exemples sont annotés manuellement. En s'appuyant sur ces deux heuristiques, Yarowsky (1995) étend son corpus grâce au classifieur entrainé. Une évaluation de cette démarche a montré de très bons résultats (90% de précision) cependant la taille de l'expérimentation est restée limitée. Une approche différente est proposée par Mihalcea *et al.* (2004) combinant deux classifieurs pour apprendre des traits différents. Un troisième classifieur combine alors les deux types d'information collectés. Certains travaux ont aussi

proposé d'améliorer ce principe de bootstrapping en utilisant le Web comme source d'exemples (Agirre *et al.*, 2001; Mihalcea, 2002).

Enfin, parmi les nombreuses approches mixtes, des travaux comme ceux de Buitelaar *et al.* (2006) ou Gliozzo et Magnini (2004) proposent de désambiguïser les sens grâce au thème d'un texte. Ces travaux combinent l'utilisation d'une ressource comme WordNet et un corpus. Mais ces méthodes sont confrontées au problème déjà pointé par Yarowsky (1992). Les noms répondent mieux que les verbes à cette technique de désambiguïsation et malgré un taux de précision bon (79%), le rappel reste bas (35%) (Gliozzo et Magnini, 2004). Ce type de résultat questionne l'intérêt de combiner ressource structurée et corpus, les défauts inhérents à une ressource comme WordNet persistent.

2.1.2 La question de l'évaluation des tâches de désambiguïsation

La question de l'évaluation s'est posée pour toutes les méthodes de désambiguïsation développées, autant à partir de dictionnaires que de corpus. Nous avons décidé de développer deux points qui nous semblent importants pour mieux comprendre dans quelle mesure ces méthodes développées pour désambiguïser peuvent ou non être appliquées en RI. Tout d'abord, nous questionnons la tâche de désambiguïsation en elle-même, sa faisabilité ; puis nous revenons sur le problème de l'évaluation d'une telle tâche.

La désambiguïsation automatique est une tâche qualifiée « d'intermédiaire » (Ide et Véronis, 1998), qui n'est donc pas un résultat en elle-même. De plus, en confrontant la désambiguïsation automatique à la nécessité de l'évaluation, s'est rapidement posé le problème de la faisabilité de la tâche. Il est difficile de réunir des conditions satisfaisantes pour évaluer la désambiguïsation, et il apparaît également que c'est une tâche difficile pour les humains. En effet le degré d'accord sur la désambiguïsation « manuelle » est très variable selon les études et le contexte (Ahlswede et Lorand, 1993; Ahlswede, 1995; Véronis, 2001).

Véronis (2001) réalise une expérience pour évaluer la capacité du jugement humain à produire un diagnostic sur la polysémie. Les résultats montrent un fort désaccord, qui varie selon les catégories morpho-syntaxiques annotées (nom, adjectif ou verbe). Dans cette expérience, il confronte également les sujets à des définitions de dictionnaire, qui s'avèrent des sources de confusion. La cause de cette confusion vient principalement du fait que les dictionnaires, jusqu'à une période récente, comportaient peu d'indices de surfaces telles que des collocations ou des coocurrences. En effet, les sujets ne peuvent réutiliser

les définitions du dictionnaire en situation lors d'un diagnostic de polysémie en contexte.

Il est donc difficile de comparer les études réalisées entre elles, et de voir quelle méthode a un réel intérêt. Ide et Véronis (1998) soulignent même le caractère artificiel et limité des évaluations et par conséquent des travaux de désambiguïsation. Ils sont en effet réalisés la plupart du temps sur des mots ayant des sens fréquents et en limitant les évaluations à une seule catégorie morpho-syntaxique, les noms. Ide et Véronis (1998) remarquent également que les seuls travaux réalisés « in vivo » donnent des résultats ambigus et mitigés. Ces premières évaluations sont des travaux réalisés en RI, en l'occurrence, Krovetz et Croft (1992) et Schütze et Pedersen (1995).

Malgré ces constats, la communauté a créé des campagnes d'évaluation basées sur des expériences « in vitro » basées sur une ressource de référence. A l'issue d'un workshop intitulé Siglex (*Special Interest Group on the Lexicon*) en 1997, il a été décidé de constituer une campagne d'évaluation dans le même esprit que celles réalisées dans le domaine de l'extraction d'information (MUC). La première campagne a eu lieu en 1998, sous le nom de *Senseval* [1], et elle a été suivie, à ce jour, de 5 éditions (2001, 2004, 2007, 2010 et 2013).

La première édition de Senseval avait pour volonté d'unifier les efforts, dans le but de constituer des ressources et d'étiqueter des corpus (Ide et Véronis, 1998). La tâche de prédilection est celle d'attribution de sens, lors de la première édition se fut le dictionnaire H, remplacé dès la deuxième édition par WordNet. Avec le temps, les campagnes ont pris de l'ampleur en multipliant les tâches d'évaluation. Ainsi, les éditions de 2007 et 2010 ont comporté 18 tâches, celles-ci étant variées et de granularité assez différentes : tâche d'extraction automatique de mots-clés dans des articles scientifques, désambiguïsation lexicale en japonais ou encore résolution de co-référence dans plusieurs langues par exemple pour Semeval 5.

Navigli (2009) estime que l'objectif de comparer les systèmes entre eux n'est pas atteint, en particulier entre les différentes campagnes ; ainsi, estimer la progression des performances est très difficile. Les conditions ont en effet beaucoup changé dans le temps, par exemple les dictionnaires de référence n'ont pas été les mêmes au cours des campagnes.

Au cours du temps, les tâches se sont multipliées et diversifiées. Le multilinguisme s'est aussi développé, même si l'anglais reste la langue « reine ». Les conditions proposées sont éloignées de la RI. Des campagnes spécifiques au

1. http://www.senseval.org/

domaine de la RI, les campagnes TREC, ont été créées dès 1992 comme nous le verrons en 2.2.2.2.

2.2 Résoudre l'ambiguïté en RI

Les premiers travaux « historiques » ont cherché à identifier les différents sens que pouvait porter une requête, et à estimer l'impact de l'ambiguïté de certaines requêtes sur les performances du système de recherche. Les travaux de Weiss (1973) ont été les premiers à essayer de mesurer l'impact de l'utilisation d'un désambiguïsateur pour améliorer la représentation d'un document en RI. Le traitement de l'ambiguïté est considéré comme un enjeu important pour l'amélioration des performances d'un système : l'apport d'une phase de désambiguïsation a été démontré à plusieurs reprises, par exemple par (Schütze et Pedersen, 1995) ou plus récemment (Stokoe *et al.*, 2003). De fait, de nombreux travaux ont été consacrés à cette question depuis les années 1990. Les solutions proposées pour la résoudre se sont d'abord focalisées sur le traitement de l'ambiguïté lexicale par recours à des dictionnaires ou, en recherche d'information multilingue (Krovetz et Croft, 1992; Sanderson, 2000; Stokoe, 2005). Cependant, le bénifice pour les système de recherche d'information est conditionné par les performances des systèmes de désambiguïsation. Peu à peu, les travaux sur ce sujet ont déplacé le problème de l'ambiguïté du plan de la désambiguïsation à celui de la résolution de l'ambiguïté.

Les techniques utilisées pour résoudre les problèmes d'ambiguïté portés par une requête sont de deux types : les méthodes qui choisissent d'agir sur les requêtes en elles-mêmes, et les méthodes qui se concentrent sur les résultats de la recherche d'information. Les techniques qui se concentrent sur la requête sont elles-mêmes de différentes sortes. Nous allons présenter ici les techniques héritées de la désambiguïsation lexicale et appliquées aux requêtes, puis des méthodes plus récentes qui tentent d'apporter le contexte manquant en utilisant l'expansion de requêtes ou de mesurer le niveau de clarté d'une requête. Les techniques de désambiguïsation s'appuyaient sur la présence d'un contexte riche induit par des requêtes de plusieurs mots.

Nous abordons ensuite les méthodes qui se focalisent sur les résultats obtenus en réponse à une recherche d'information. Ces méthodes sont avant tout de deux sortes : le clustering des résultats et la réorganisation de ceux-ci. Enfin, nous évoquons la façon dont ces résultats de recherche peuvent offrir des solutions pour résoudre les problèmes d'ambiguïté en RI (Hearst, 2009).

2.2.1 Les indices pour désambiguïser en RI

Nous proposons de considérer des indices importants listés par Audibert (2003) pour la désambiguïsation lexicale et de les confronter à la situation de la RI afin de voir lesquels sont opérationnels. Nous rappelons que les requêtes sont seulement composées de quelques mots juxtaposés.

Selon Audibert (2003), on peut dénombrer au moins 8 indices (que nous avons regroupé en 5 points) qui permettent d'améliorer la désambiguïsation lexicale :

- l'étiquetage morphosyntaxique et les indices syntaxiques (dont les contraintes de sélection) : l'étiquetage morphosyntaxique permet de lever l'ambiguïté sur la catégorie grammaticale des mots. C'est une opération difficile à réaliser sur les requêtes de un ou deux mots, les erreurs étant très fréquentes lorsque le contexte est absent. Les indices syntaxiques sont également importants pour la désambiguïsation. Toutefois, les informations syntaxiques sont rares dans les requêtes, où les marques syntaxiques disparaissent au profit de mots juxtaposés.
- les collocations et les cooccurrences : les collocations sont des mots qui entretiennent une relation privilégiée avec le mot-cible à désambiguïser. Tout comme l'indice des cooccurrences, les collocations s'appuient sur un contexte autour du mot à désambiguïser, contexte souvent absent en RI.
- les organisations en taxinomies et les associations thématiques : les taxinomies relient les différentes lexèmes d'un dictionnaire, et permettent ainsi d'améliorer la réussite de la désambiguïsation. Cependant, ce type de traitement ne permet pas de traiter les noms propres très présents dans les requêtes (cf. chapitre 1). Les associations thématiques sont également un indice intéressant, puisqu'elles donnent la possibilité de relier des mots comme *garçon* au sens de *garçon de café* et *table*.
- l'information sur le thème du texte : elle permet de savoir quel est le sujet du texte où se trouve le mot à désambiguïser. Ainsi, si un document est un texte juridique, le mot *avocat* aura plus de chance de prendre le sens d'*homme de loi*, plutôt que de *fruit*. C'est un critère difficilement applicable à une requête, par contre il est possible de considérer le thème des documents où apparaît la requête si cette information est disponible.
- la fréquence des sens : cet indice est utile lorsqu'aucune autre information n'est disponible.

On voit donc qu'un certain nombre de points forts pour la désambiguïsation lexicale sont absents ou inaccessibles en RI.

2.2.2 Résoudre l'ambiguïté : l'action sur la requête

Les premiers travaux hérités de la désambiguïsation lexicale ont travaillé sur les requêtes des utilisateurs (Krovetz et Croft, 1992; Voorhees, 1993; Schütze et Pedersen, 1995; Gonzalo *et al.*, 1998). Comme nous allons le voir, ces travaux se sont principalement heurtés à l'absence de contexte. Une deuxième vague de travaux a donc essayé de répondre à ce problème en utilisant l'expansion de requêtes. Cependant, cette méthode intéressante pose de nombreuses questions techniques. Enfin, nous présentons une troisième type de méthode qui s'attache à mesurer le degré d'ambiguïté d'une requête.

2.2.2.1 Les techniques héritées de la désambiguïsation lexicale

Les travaux cherchant à désambiguïser les requêtes en appliquant des méthodes héritées de la désambiguïsation lexicale s'appuient principalement sur la recherche de collocations ou de cooccurrences à l'intérieur de la requête.

Les premiers à utiliser ce type de méthode sur les requêtes sont Krovetz et Croft (1992). Ils travaillent à partir de deux corpus de requêtes : CACM (communications de ACM - 64 requêtes) et Times (journal - 83 requêtes). Ils montrent avant tout l'importance du contexte dans ce type d'approche. En effet, ils aboutissent à la conclusion qu'il existe un « effet » des collocations, ce qui permet de limiter l'ambiguïté présente dans les requêtes. Ainsi, le mot *bank*, seul est ambigu. Mais s'il s'insère dans une requête comme *bank economic financial monetary fiscal*, l'effet de collocation peut être exploité.

Krovetz et Croft (1992) montrent ainsi que la désambiguïsation a un effet positif sur la précision du système, même s'ils attribuent cette réussite à l'utilisation du stemming. En effet, la présence des collocations les pousse à conclure que les situations où la désambiguïsation peut être utile sont au final rares. Toutefois, leur étude se base sur des requêtes assez longues, les requêtes ont une longueur moyenne 9,46 mots (CACM) et 7,44 mots (Times). Or la longueur d'une requête est un facteur important selon Sanderson (1994).

Stokoe *et al.* (2003) vont combiner les différents atouts mis en avant dans les précédents travaux de désambiguïsation en RI : les cooccurrences (Schütze et Pedersen, 1995), les collocations (Krovetz et Croft, 1992) et les fréquences des sens qui sont repérés grâce à WordNet. L'évaluation de cette approche montre une amélioration significative par rapport à la *baseline* choisie par Stokoe *et al.* (2003), correspondant à un TF*IDF. En effet, les erreurs créées par les systèmes de désambiguïsation ont été réduites, mais on observe que les noms propres ne sont pas pris en compte dans l'expérience. Les facteurs propres à

la RI comme des requêtes et des documents courts, contenant de nombreuses entités nommées, ne sont pas pris en compte, les systèmes sont en échec et l'ambiguïté reste un problème (Sanderson, 2000).

2.2.2.2 La question de l'évaluation en RI

Ces travaux utilisant les méthodes de désambiguïsation lexicale posent également la question des données utilisées car ils reposent sur une longue tradition d'expérimentations réalisées avec des requêtes longues composées de plusieurs mots (Spärck-Jones *et al.*, 2007). En effet, comme nous l'avons vu, Krovetz et Croft (1992) ont utilisé des requêtes relativement longues (entre 7 et 9 mots). D'autres comme Gonzalo *et al.* (1998) ont reconstruit des requêtes qui s'apparentent à des résumés de textes. Enfin, les travaux plus récents comme ceux de Stokoe *et al.* (2003) s'appuient sur des données issues de TREC [2] (Text REtrieval Conference). C'est une conférence créée en 1992, organisée de manière annuelle, afin de donner une infrastructure et un cadre pour mener des évaluations autour de la recherche d'information. Les données produites dans le cadre de ces évaluations sont largement reprises par la communauté. Une requête TREC (cf. figure 2.1) se présente de la manière suivante :

- un champ *num* ;
- un champ *title* qui contient la requête ;
- un champ *desc* qui contient la tâche à réaliser ;
- un champ *narrative* qui précise quels documents doivent être considérés comme résultats à cette requête cela permet d'évaluer la pertinence des documents retournés.

<num> Number : 468
<title> incandescent light bulb
<desc> Description : Find documents that address the history of the incandescent light bulb.
<narr> Narrative : A relevant document must provide information on who worked on the development of the incandescent light bulb. Relevant documents should include locations and dates of the development efforts. Documents that discuss unsuccessful development attempts and non-commercial use of incandescent light bulbs are considered relevant.

FIGURE 2.1 : Exemple d'une requête de TREC 9 Web Track emprunté à Stokoe *et al.* (2003)

2. Text REtrieval Conference http://trec.nist.gov/

Certaines critiques ont pu être émises à l'égard de ce type de données, particulièrement pour la campagne d'évaluation CLEF (Cross-Language Evaluation Forum) qui propose des tâches de RI et questions-réponses multilingues sur le même modèle que TREC. Mur (2006) mais également van der Plas (2008) soulèvent le caractère parfois très artificiel des requêtes proposées, certaines étaient identiques à une phrase présente dans un document de la collection.

De plus, les méthodes développées dans ces conditions sont difficilement applicables dans un cadre applicatif.

2.2.2.3 L'expansion de la requête : une solution pour le manque de contexte

Une autre des solutions apportées pour résoudre l'ambiguïté des requêtes est l'expansion de la requête. En effet, comme le souligne Sanderson (1994, 1997), la longueur des requêtes est un élément crucial. Étendre une requête courte permet de reconstituer un contexte linguistique absent. L'expansion permet aussi de s'intéresser à un autre type d'ambiguïté, proche de celui décrit en 1.2.2.

Allan et Raghavan (2002) ont proposé une méthode de ce type. Leur méthode d'expansion est orientée de manière à obtenir les différents sens d'une requête ambiguë en contexte. En effet, ils construisent des patrons qui vont ensuite permettre de générer des requêtes étendues contenant la requête cible comme par exemple pour *party* :

JJ party : NN varieties of a *party*
NP party : NP names of a *party*
NN IN party : NN NNS things done with a *party*
VB party : NN NN things done to a *party*

Ces requêtes étendues sont alors soumises au moteur de recherche. Cette technique est intéressante parce qu'elle fait intervenir une prise en compte de l'ambiguïté en contexte, elle permet en effet un bon niveau de clarification (41% d'amélioration sur les requêtes TREC et 25% sur les requêtes Web). Toutefois, cette méthode est fortement liée au corpus qui sert à générer les « patrons ». En effet, il est nécessaire d'avoir un corpus statique et étiqueté syntaxiquement. Cela rend l'expansion de requêtes coûteuse en traitement en amont de la recherche d'information.

Santos *et al.* (2010b) considèrent qu'une requête ambiguë peut être décomposée en plusieurs « sous-requêtes » ce qui permet de considérer la diversité portée par une requête ambiguë. Le corpus de requêtes ambiguës étudiées provient de TREC 2009 Web track. Des balises subtopic signalent les différentes

facettes attribuées à la requête en question (topic) comme on peut le voir dans la figure 2.2. On voit alors que les différents subtopic font écho à ce que l'on peut trouver dans une base de documents, tel qu'une photo du journal Time ou une information sur la mère de Barack Obama.

<topic number="1" type="faceted">
<query>obama family tree</query>
<description> Find information on President Barack Obama's family history, including genealogy, national origins, places and dates of birth, etc. </description>
<subtopic number="1" type="nav"> Find the TIME magazine photo essay "Barack Obama's Family Tree". </subtopic>
<subtopic number="2" type="inf"> Where did Barack Obama's parents and grandparents come from ?</subtopic>
<subtopic number="3" type="inf"> Find biographical information on Barack Obama's mother. </subtopic>
</topic>

FIGURE 2.2 : Exemple de requêtes TREC 2009 Web Track emprunté à Santos *et al.* (2010b)

Ces auteurs cherchent à faire émerger les différents aspects portée par une requête grâce à un algorithme de réordonnancement des résultats. Le réordonnancement est basé sur deux critères : la pertinence du document et la diversité du document. Ce critère de diversité du document est composé à partir du calcul du nombre de documents ramenés par chaque sous-requête dans Google, d'une mesure de couverture du document et d'une mesure concernant la récence du document. Le calcul du nombre de documents ramenés dans Google est censé traduire l'intérêt des utilisateurs pour chaque sous-requête.

Les sous-requêtes sont des reformulations provenant de moteurs commerciaux. En effet, les requêtes reformulées peuvent être considérées comme des « sous-requêtes » d'une requête ambiguë. Santos *et al.* (2010b) ont collecté deux types de sous-requêtes : des requêtes liées aux résultats ramenés par la première requête et des requêtes suggérées sous la requête initiale par les systèmes.

Si Santos *et al.* (2010b) concluent que les reformulations sont des représentations intéressantes de la diversité contenue dans la requête initiale, il semblerait que les critères choisis pour discriminer les différentes sous-requêtes gagneraient à être complétés par d'autres critères comme le type d'information

porté par la sous-requête (navigationnelle ou informationnelle) [3].

2.2.2.4 Les mesures évaluant la clarté de la requête

Certains travaux ont appréhendé le problème de l'ambiguïté des requêtes différemment. Au lieu de chercher à identifier les différents sens que porte une requête, ils calculent un degré d'ambiguïté (Cronen-Townsend et Croft, 2002; Allan et Raghavan, 2002). Le traitement de l'ambiguïté est alors envisagé comme un diagnostic, et non plus comme une tâche de désambiguïsation.

Ce diagnostic d'ambiguïté se calcule principalement grâce à une mesure appelée « mesure de clarté » (Cronen-Townsend et Croft, 2002). Cela signifie que moins une requête sera claire, plus elle sera ambiguë. Cronen-Townsend et Croft (2002) appliquent la mesure de *Kullback-Leilber divergence (Cover et Thomas, 2012)* (ou d'entropie relative) au contexte de RI. Elle dépend fortement de la collection de documents sur laquelle la requête va être projetée. En effet, les auteurs signalent que lorsqu'une requête occurre dans peu de documents de la collection étudiée, le degré de clarté augmente. Par exemple, la requête *jobs* face à une collection TREC choisie par Cronen-Townsend et Croft (2002) a une mesure de clarté de 0,29. Alors que la requête *steve jobs* est à 0,53 et *denim textiles jobs* à 2,16. Ces exemples montrent que cette mesure est graduelle.

Cette méthode peut également s'appliquer non pas sur la collection de documents mais sur les résultats cliqués par les utilisateurs (c'est-à-dire les documents qui ont été consultés par les utilisateurs suite à une requête, *cf.* chapitre 3). Dou *et al.* (2007) calculent alors une mesure appelée « click entropy ». Cette mesure repose sur l'hypothèse que si les utilisateurs cliquent sur un grand nombre d'URL différentes proposées en résultats, cela signale une diversité de réponses possibles à la requête et par, là même, une ambiguïté potentielle. Ainsi, si tous les utilisateurs cliquent sur la même page pour une requête q, alors la mesure d'entropie des clics pour la requête q sera égale à zéro. Ce qui signifiera que la requête n'est pas ambiguë. Ce type de démarche est à mettre en lien avec la volonté de personnaliser les résultats de recherche, c'est-à-dire proposer des résultats adaptés à l'utilisateur, qui tiennent compte de ses spécificités.

2.2.3 Révéler l'ambiguïté : l'action sur les résultats

Différents travaux choisissent d'agir directement sur les résultats d'une recherche d'information pour agir sur les problèmes d'ambiguïté des requêtes.

3. Notion développée dans le chapitre 3

En effet, l'ambiguïté d'une requête est révélée par les résultats auxquels elle permet d'accéder. L'hypothèse est donc que la modification du regroupement ou de l'ordonnancement des résultats permet de mettre au jour les différentes dimensions du sens d'une requête et d'améliorer la satisfaction de l'utilisateur. Pour agir sur le regroupement des résultats, de nombreux travaux mettent en place du clustering de résultats. Nous allons donc voir dans un premier temps ces méthodes et leurs avantages. Puis dans un second temps, nous nous focalisons sur les méthodes qui touchent au réordonnancement des résultats.

2.2.3.1 Le clustering de résultats

Face aux difficultés rencontrées par les méthodes de désambiguïsation lexicale appliquées la RI, des travaux alternatifs se sont développés pour devenir aujourd'hui majoritaires. La résolution de l'ambiguïté est abordée grâce à des procédures de clusterisation des textes, qui visent à faire émerger les différents emplois des mots de la requête représentés dans la base documentaire interrogée, reprenant les principes développés par les approches de désambiguïsation sur corpus (cf. 2.1.1.2). Le *clustering* de documents consiste à regrouper des données brutes, sans autre information externe.

Le clustering de résultats présente un certain nombre de défis par rapport à une tâche de classification de textes comme le signalent Carpineto *et al.* (2009). Le principal défi est la création de « labels » qui sont pertinents pour l'utilisateur et doivent rendre compte des différents sens, en décrivant les différences entre des clusters génériques et d'autres plus spécifiques. Les labels doivent exprimer la granularité variable des clusters. Il existe deux grandes familles d'algorithmes utilisés pour le clustering de résultats (Carpineto *et al.*, 2009) :

– les *data-centric algorithms* ;
– les *description-centric algorithms*.

Les algorithmes dits *data-centric* sont des algorithmes de clustering de données « conventionnels » appliqués aux clustering des résultats de recherche. Ils incluent souvent des options de présentation spécifiques pour la représentation textuelle des clusters et le résultat final soumis aux utilisateurs, grâce à du clustering hiérarchique (Maarek *et al.*, 2000) ou en utilisant des liens informatifs comme les liens hypertextes ou thématiques (Zhang *et al.*, 2008). Ce type d'approche se trouve en difficulté lorsqu'il s'agit de générer des labels appropriés pour chaque cluster. L'étiquetage consiste à utiliser comme label les termes les plus fréquents de chaque cluster. Le résultat de cette labellisation n'est pas forcément compréhensible et interprétable pour l'utilisateur (Navarro *et al.*, 2011).

Le second type d'algorithmes dits *description-centric* sont spécifiquement conçus pour regrouper des résultats de recherche (Zamir *et al.*, 1997; Zamir et Etzioni, 1998; Osinski *et al.*, 2004). Ils prennent donc en compte la qualité du clustering et les descriptifs des clusters. Ces algorithmes servent également à extraire des labels cohérents des documents (Crabtree *et al.*, 2005; Bernardini *et al.*, 2009). Ces méthodes peuvent soit construire des clusters de documents qui sont ensuite étiquetés, soit construire des clusters d'étiquettes auxquels elles attribuent des documents (Navarro *et al.*, 2011). On retrouve ce type de méthodologie chez les moteurs de recherche intégrant un clustering sur les résultats comme Vivisimo [4] ou Carrot² [5].

Du point de vue technique, plusieurs différences existent entre les techniques de clustering sur documents et celles sur des résultats de recherche. Le clustering des résultats est souvent effectué en ligne, il doit donc fournir ses résultats le plus rapidement possible. Le temps de traitement doit donc être plus court par rapport à un algorithme traditionnel. Le clustering de recherche doit s'adapter à des données réduites souvent limitées à une URL, un titre, un court descriptif appelé snippet. Il doit également calculer un nombre de clusters qui est inconnu à l'avance, tout comme la taille des clusters, autant de chose que les algorithmes traditionnels ne peuvent accepter.

Par ailleurs, Hearst (2009, section 8.2) relève que le clustering est utile pour clarifier une requête « vague » ou pour désambiguïser des acronymes. Cependant, le clustering est une solution difficile à déployer. Contrairement aux algorithmes de clustering qui se limitent à proposer des visualisations des résultats sous forme de listes ou de graphiques principalement destinés à des spécialistes du domaines, le clustering de résultats de recherche doit présenter ses résultats d'une manière lisible et intuitive pour les utilisateurs. Or, les documents peuvent être rattachés à plusieurs clusters, et ce qui rend leur représentation difficile dans l'interface des résultats (Hearst, 2009, 2011).

2.2.3.2 La réorganisation des résultats de recherche

C'est dans ce contexte qu'émerge la notion de « diversité », suggérant un autre type de solution pour révéler l'ambiguïté portée par les requêtes des utilisateurs (Chen et Karger, 2006). C'est une notion qui est propre au domaine des moteurs de recherche Web. Chen et Karger (2006) avancent que l'enjeu n'est pas de retrouver le plus de documents possibles répondant à une requête mais de trouver les meilleurs documents. Sont considérés comme étant les

4. http://search.vivisimo.com/
5. http://search.carrot2.org/stable/search

meilleurs documents ceux qui représentent la diversité d'une requête. Hearst (2009) pointe le fait que cette notion de diversité est importante pour les requêtes ambiguës :

> « Some Web search engines attempt to support a notion of "diversity" in the first few results displayed. This is especially important for ambiguous queries for which there are several common interpretations or meanings for a given word. » (Hearst, 2009, section 5.4)

Ce type de démarche peut soit utiliser une catégorisation des textes pour effectuer la réorganisation des résultats, soit s'employer directement à calculer une réorganisation des résultats (Zhai *et al.*, 2003; Zhai et Lafferty, 2006). Un certain nombre de services de moteurs de recherche ont utilisé des catégories pour organiser ces résultats de recherche : Yahoo !, Snap ou LookSmart (Chen et Dumais, 2000). Ces services associaient une catégorie labellisée à chaque résultat de recherche. Les résultats sont alors présentés sous la forme d'une liste des catégories en rendant visible le niveau le plus bas de la hiérarchie. Ces techniques restent tout de même majoritairement utilisées dans des domaines spécifiques (comme le médical), s'appuyant sur des thésaurus. Ces hiérarchies de catégories sont souvent constituées manuellement et demandent donc un effort initial conséquent, mais la difficulté principale reste la maintenance d'un tel système. Cependant, Dumais *et al.* (2001) ont montré que l'usage de catégories dans la présentation des résultats maximise la désambiguïsation des requêtes ambiguës, et particulièrement pour des requêtes que l'on pourrait qualifier de « larges » (cf. 1.2.2).

De nombreux travaux ont, quant à eux, essayé de catégoriser automatiquement les résultats d'une recherche pour faire émerger la diversité des requêtes. Ce type de travaux s'appuient sur le principal annuaire du Web, Open Directory Projet (ODP), aussi appelé *dmoz*[6]. Dans cette démarche, Zhang *et al.* (2005); Xue *et al.* (2008) ont étendu la classification présente dans cet annuaire aux pages web dans leur ensemble afin de pouvoir générer des résultats catégorisés. Cette approche consiste à utiliser des techniques statistiques pour apprendre un modèle basé sur un corpus de documents catégorisés. Ce modèle est ensuite appliqué aux nouveaux documents (sans catégorie déterminée) afin de déterminer leur catégorie d'appartenance. Les informations ainsi ajoutées aux documents servent ensuite à renvoyer des résultats retravaillés.

Nous avons également vu que les travaux de Santos *et al.* (2010b) utilisent le réordonnancement des résultats pour résoudre l'ambiguïté des requêtes (cf.

6. http://blog.dmoz.org/

2.2.2.3). Leurs travaux s'appuient à la fois sur les requêtes et les résultats pour arriver à faire émerger la diversité.

Zhang *et al.* (2005) proposent alors la combinaison de deux principes : la *diversité* mesure la variété des sujets dans un groupe de documents (calculée grâce aux catégories de l'ODP) et l'*intensité* de l'information mesure combien un même document contient de sujets différents. Cette combinaison donne lieu à un nouvel algorithme, *Affinity Ranking*, qui réorganise le classement des documents proposés comme résultats de recherche. Les résultats des expériences sur des requêtes de 1 à 3 mots montrent que cette méthode peut améliorer la diversité et la richesse de l'information des 10 premiers résultats de recherche sans perdre en pertinence. C'est donc une solution intéressante pour le cas des requêtes courtes, touchées par de l'ambiguïté.

Hearst (2009) souligne également que les systèmes à partir de catégories ou de facettes sont plus faciles à utiliser que les systèmes à base de clustering pour les utilisateurs (Käki, 2005). En effet, les utilisateurs préfèrent les hiérarchies qui ont du sens et dont la granularité est uniforme (Hearst, 2011). Il est plus facile de naviguer dans un système à catégories, même si les catégories utilisées peuvent parfois paraître trop générales ou peu en rapport avec les sens des documents. Ces résultats montrent la difficulté qu'il y a à combiner intérêt de l'utilisateur et la facilité technique sur cette question d'ambiguïté des requêtes.

2.3 Conclusion

La désambiguïsation lexicale reste un problème difficile, qui l'est autant pour un système automatique que pour un humain (Ahlswede et Lorand, 1993; Ahlswede, 1995; Véronis, 2001). Ces constatations permettent de relativiser l'échec de certains systèmes, en particulier face à la désambiguïsation des requêtes où le contexte est très réduit voire absent.

D'autre part, les solutions apportées par la désambiguïsation automatique sont peu adaptées pour traiter le problème de l'ambiguïté des requêtes. Elles ont été en effet envisagées dans un contexte « artificiel » (Ide et Véronis, 1998) et n'ont pu être transposées dans un contexte « réel » de RI. Les solutions envisagées par la suite confirment la difficulté de la tâche.

Par ailleurs, la question de la résolution de l'ambiguïté place l'utilisateur au centre du raisonnement. Il ne s'agit plus de désambiguïser une requête pour un système, mais de clarifier des résultats pour un utilisateur.

En effet, qu'il s'agisse des approches traitant des requêtes directement ou celles se focalisant sur les résultats de recherche, toutes se heurtent à la difficulté de traiter l'ensemble des types de requêtes ambiguës. Le clustering de résultats comme le souligne Hearst (2009) serait plus adapté pour traiter des ambiguïtés qui touchent les acronymes (homonymie). Cependant, comme nous l'avons vu dans le chapitre 1, l'ambiguïté des requêtes ne se limite pas à ce type d'ambiguïté. Ainsi, la mise en place de catégories réorganisant les résultats résoudrait mieux l'ambiguïté provenant des requêtes « larges ».

Enfin, l'ensemble des travaux souligne que les éléments du contexte de la recherche d'information sont indispensables pour repérer l'ambiguïté (reformulations, clics des utilisateurs, etc.). L'état de l'art confirme donc qu'il y a un intérêt certain à la prise en compte d'éléments du contexte de la recherche d'information pour améliorer les performances des systèmes et pour traiter le problème de l'ambiguïté.

Chapitre 3

La recherche d'information et l'apport du contexte

Nous venons de voir dans le chapitre 2 que les méthodes de désambiguïsation étaient inadaptées pour traiter le problème de l'ambiguïté des requêtes. Les travaux examinés soulignent que les éléments du contexte de la recherche d'information sont indispensables pour repérer l'ambiguïté. Par conséquent, nous allons voir dans ce chapitre ce que sont ces éléments contextuels et dans quelle mesure ils peuvent apporter des solutions. Cette démarche nous amène à nous intéresser à la RI contextuelle dont l'objectif est de mieux répondre au besoin d'information des utilisateurs. Pour cela, la RI contextuelle cherche à combiner les technologies de recherche avec des connaissances sur les requêtes et sur le contexte utilisateur (Allan *et al.*, 2003). Pour aborder ces questions, nous allons introduire dans un premier temps les étapes basiques qui composent la recherche d'information dite « traditionnelle ».

Puis dans un deuxième temps, nous abordons la question du contexte en RI. Nous voyons que les éléments du contexte concernent les intentions de l'utilisateur, l'environnement de l'utilisateur et le système lui-même (Mothe, 2011). Ce constat nous amène à expliquer comment peut-on le prendre et comment sa prise en compte peut aider à l'amélioration des performances en RI.

3.1 La recherche d'information

La recherche d'information s'intéresse à « l'acquisition, l'organisation, le stockage, la recherche et la sélection d'information » (Pinel-Sauvagnat, 2005). Ce sont les systèmes de recherche d'information qui permettent aux utilisateurs de retrouver un document correspondant à une requête exprimée en langage

naturel. L'accès à l'information peut se faire dans un contexte varié : recherche sur le web, recherche dans un domaine spécialisé. La recherche ne se limite plus à l'interrogation d'une base par le biais de notices bibliographiques classées par mot-clés, titres ou auteurs. Désormais, la recherche d'information permet d'accéder à l'information contenue dans les documents, en ne s'appuyant plus exclusivement sur des caractéristiques exogènes au document. Nous allons présenter le processus sur lequel reposent les systèmes de recherche d'information, qui est appelé *processus en U* (Salton et McGill, 1986), et dont nous proposons une vue schématique en 3.1 (schéma emprunté à Chevalier (2011)).

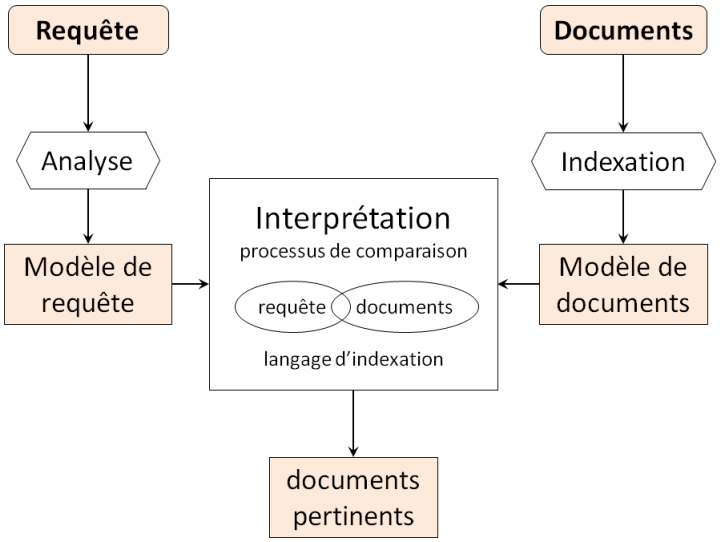


FIGURE 3.1 : Le processus en « U » de la recherche d'information par Chevalier (2011)

Selon Croft (1993), un système de RI met en œuvre trois processus : la représentation du contenu des documents de la collection (branche « documents » sur la figure 3.1), la représentation du besoin d'information de l'utilisateur (branche « requête »), et la comparaison de ces deux représentations figurée par le processus d'interprétation. Sur la figure 3.1, sont distingués par un code couleur les processus impliqués dans la recherche d'information et les données utilisées par ces processus. Les premiers sont en blanc, les seconds en rose.

Un processus de RI, comme figuré sur la figure 3.1, se décline en quatre grandes étapes :

1. Formulation des besoins par l'utilisateur

2. Indexation de la collection de documents à disposition
3. Appariement des requêtes et des documents
4. Présentation des résultats à l'utilisateur

Nous allons présenter ces étapes d'un processus classique de RI, en nous focalisant sur les éléments qui vont être déterminants pour la suite de nos travaux.

3.1.1 Le processus de formulation d'une requête

La formulation d'une requête découle d'un processus de recherche enclenché par l'utilisateur. En effet, celui-ci va être confronté à plusieurs tâches cognitives qui vont aboutir à la soumission d'une requête à un système de recherche, sous forme de langage naturel. Le processus de formulation s'inscrit dans le cadre d'un processus plus large décrivant une interaction entre un utilisateur et un système. De nombreux auteurs ont décrit ce processus de recherche d'information. Hearst (2009) propose une synthèse de ces travaux.

La formulation d'une requête est précédée dans la plupart des modèles par deux étapes : l'identification du problème et l'articulation de ce problème aux besoins informationnels (Sutcliffe et Ennis, 1998; Broder, 2002). Marchionini et White (2008) proposent un modèle plus détaillé où une étape d'acceptation de la tâche par l'utilisateur est insérée entre ces deux étapes (voir la figure 3.2 pour une description de l'ensemble des étapes). Le processus de formulation résulte donc d'une interaction engagée entre l'utilisateur et le système.

Marchionini et White (2008) détaillent le processus de formulation de la requête en deux phases : la formulation du problème et l'expression d'un besoin d'information au système de recherche. Cette dissociation du processus de formulation est également mise en avant chez Mizzaro (1998) qui distingue lui aussi la phase d'expression.

L'expression de la requête peut être confrontée à deux problèmes : le « label effect » et la variabilité du vocabulaire (Taylor, 1968; Ingwersen, 1992; Belkin *et al.*, 1982a,b). Le *label effect* a été défini en 1968 par Taylor lors d'études menées sur les processus de recherche d'information opérés par des documentalistes : « The label effect is thus a manifestation of the conceptual distance δ between underlying need and actual request formulation » (Ingwersen, 1992). Il décrit le fait qu'une expression sous forme de label ou de mot-clé a tendance à ne pas formuler clairement tous les aspects du besoin informationel (Ingwersen, 1992).

Le *label effect* est un problème dû au contenu linguistique de la requête. On constate la non-correspondance entre les termes utilisés dans la requête et

ceux utilisés dans les documents. La diversité linguistique n'est pas le seul problème, les nombreux termes ambigus et les phénomènes de synonymie peuvent perturber l'expression du besoin informationnel.

L'interaction entre l'utilisateur et le système ne se termine pas avec la soumission de la requête mais avec une phase de vérification des résultats (Shneiderman *et al.*, 1997), suivie d'une reformulation du problème et de son expression si cela est nécessaire avant d'utiliser les résultats (Marchionini et White, 2008). Ces étapes peuvent être une source d'information comme nous le verrons dans la section 3.2.3.2 de ce chapitre.

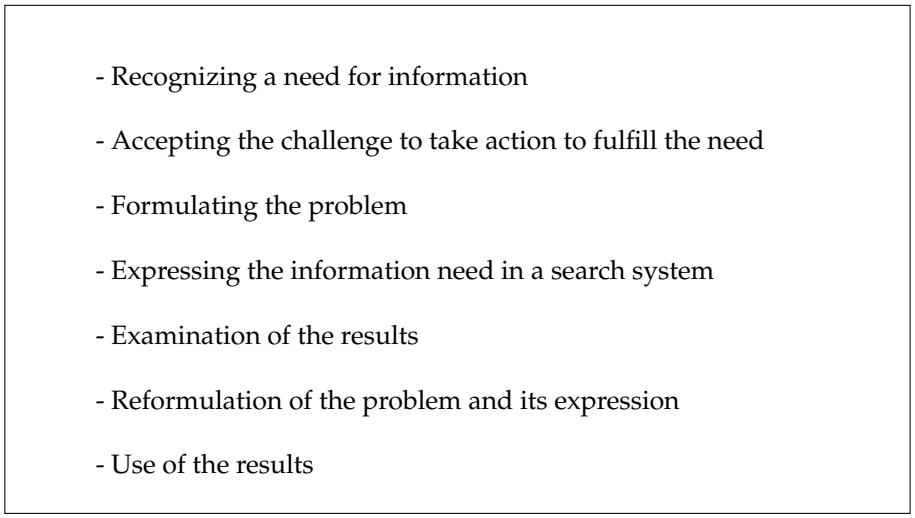


FIGURE 3.2 : Description du processus de recherche par Marchionini et White (2008) dans Hearst (2009)

3.1.2 Les modèles classiques de recherche d'information

Un système de RI réalise une indexation de la collection de documents dans laquelle l'utilisateur cherche une information. Cette phase d'indexation permet d'identifier les caractéristiques les plus importantes des documents, en ne gardant que les unités lexicales susceptibles d'être de bons descripteurs et en excluant les « stop-words » tels que les prépositions ou les déterminants. Cette phase d'indexation est suivie par une phase d'appariement entre la requête et les documents pertinents, cet appariement peut s'opérer de différentes manières (Manning *et al.*, 2008).

Nous présentons ici les trois modèles « classiques » qui sont utilisés en recherche d'information. Les modèles sont présentés dans l'ordre de leur ap-

parition historique, ordre qui figure également leur niveau de complexité. Le modèle booléen est en effet le plus simple et un des premiers à avoir été utilisé. Le deuxième modèle est le modèle vectoriel, modèle de RI statistique. Enfin, nous décrivons rapidement le troisième modèle utilisé en RI qui est le modèle probabiliste.

3.1.2.1 Le modèle booléen

C'est le modèle de RI le plus simple. Pour ce modèle, soit un document est pertinent, soit il est non pertinent, pour une requête donnée (Salton, 1968). Pour qu'il y ait pertinence, le document doit correspondre parfaitement aux conditions imposées par la requête. Dans ce modèle, une requête se définit de la manière suivante :

> « Une requête q est composée de termes liés par les trois connecteurs logiques ET, OU, NON. » (Pinel-Sauvagnat, 2005)

C'est donc seulement quand il y a correspondance totale entre les conditions intégrées à la requête (mots présents et opérateurs logiques) qu'un document peut être considéré comme pertinent (Salton *et al.*, 1983). Toutefois, si ce modèle est simple du point de vue du processus de recherche, la formulation d'une requête est compliquée à cause des connecteurs logiques que doit manipuler l'utilisateur.

3.1.2.2 Le modèle vectoriel

C'est un modèle statistique. Il a également été proposé par Salton et McGill (1986). L'utilisation d'un modèle statistique en RI permet deux choses selon Pinel-Sauvagnat (2005) :

- une caractérisation quantitative des documents et des termes qui les contiennent ;
- une mesure le degré de pertinence existant entre la requête et les documents.

Chaque document est représenté sous la forme d'un vecteur qui contient les termes pondérés dans le document, la requête est également représentée sous forme de vecteur. C'est grâce au calcul de cosinus de l'angle entre les vecteurs du document et de la requête que l'on obtient un degré de similarité (Manning *et al.*, 2008).

Cette représentation permet de prendre en compte la pondération des termes dans le choix des documents, et ainsi d'inclure des documents correspondant « approximativement » à la requête. Cela permet de ne pas se limiter à une

correspondante stricte comme avec le modèle booléen. Comme le souligne Pinel-Sauvagnat (2005), l'avantage principal de ce modèle est de permettre le tri des documents en fonction de leur similarité avec la requête, son avantage secondaire est la simplicité, ce qui en faisait le modèle plus utilisé en RI pour les premiers moteurs commerciaux.

3.1.2.3 Les modèles probabilistes

Le principe d'un modèle probabiliste est d'ordonner les résultats en réponse à une requête selon leur probabilité de pertinence et non selon leur similarité comme dans le cas du modèle vectoriel. Ce principe est appelé « principe de classement probabiliste » par Robertson (1977) (*Probability Ranking Principle* soit PRP).

Le calcul du degré de probabilité de pertinence est basé sur deux probabilités conditionnelles (Pinel-Sauvagnat, 2005) :

– la probabilité qu'un terme_i occurre dans un document_j qui soit pertinent pour la requête ;
– la probabilité qu'un terme_i occurre dans un document_j qui ne soit pas pertinent pour la requête.

Parmi les exemples de modèles RI probabilistes, on retrouve le modèle 2-Poisson de Robertson et Walker (1994) et le modèle Okapi de Walker *et al.* (1998).

3.1.3 La présentation des résultats

Comme nous l'avons vu en 3.1.1, la présentation des résultats à l'utilisateur est une étape du processus de recherche. Cette présentation vient répondre à la requête formulée par l'utilisateur. Les résultats peuvent être présentés de différentes manières, le plus souvent sous forme de listes ordonnées verticales ou sous forme de catégories. Mais le mode de présentation le plus utilisé est celui sous format de listes ordonnées verticales (Hearst, 2009).

Ce mode de présentation a pourtant des inconvénients, en particulier lorsque la requête soumise est ambiguë. En effet, comme le signalent Navarro *et al.* (2011), face à une requête ambiguë, une présentation affichant les résultats sous forme de listes n'a que trois choix possibles : une désambiguïsation tacite, une désambiguïsation profilée ou un assortiment. Prenons pour exemple, la requête *orange*. Un affichage pratiquant une désambiguïsation tacite va privilégier le sens le plus recherché, en l'occurrence l'entreprise. Cette stratégie

peut être complétée par de la désambiguïsation profilée qui va afficher un résultat en fonction du profil de l'utilisateur (cf. 3.2.3.3), comme on peut le voir sur la figure 3.3. Ces résultats affichés par Google sont personnalisés et privilégient les liens déjà cliqués (en mauve). Enfin, le choix de l'assortiment va apparaître sous la forme d'une liste de documents composés des différentes interprétations que peut prendre la requête. C'est le choix du moteur de Orange (figure 3.4) qui propose des résultats sur Orange la marque et sur la couleur orange (signalant un phénomène dangereux en météo). L'inconvénient de ce type de liste est l'impression de « panachage » des résultats présentés.

Parmi les autres choix de présentations les plus populaires, on peut citer le clustering et le filtrage par catégorie (Hearst, 2006). Nous avons vu dans le chapitre 2 que la présentation des résultats issus du clustering reste difficile à mettre en place de manière satisfaisante, contrairement à la mise en place de catégories.

Les étiquettes des catégories sont souvent déterminées manuellement, puis attribué aux documents de manière automatique (Hearst, 2006). La figure 3.5 montre un exemple de résultats classés par catégorie en réponse à la requête *orange* (Qwant). On voit que deux interprétations de la requête dominent, l'entreprise et le fruit [1].

La question de la présentation des résultats s'inscrit dans les problémantiques relevant des interfaces de recherche et d'ergonomie. La présentation des résultats doit répondre aux principes d'« usabilité » ou « utilisabilité » (Shneiderman et Plaisant, 2004). Ces principes se déclinent en cinq points qui sont définis de la manière suivante par Nielsen (2003) :

- Apprentissage : est-ce qu'il est possible pour l'utilisateur qui rencontre une interface pour la première fois d'accomplir une tâche simple ?
- Efficacité : lorsque les utilisateurs connaissent l'interface, à quelle rapidité effectuent-il une tâche ?
- Mémorisation : lorsqu'un utilisateur n'a pas utilisé l'interface depuis un certain temps, est-ce qu'il arrive à rétablir facilement ses compétences et à réutiliser l'interface ?
- Prise en compte des erreurs : combien d'erreurs font les utilisateurs ? Quelle est la gravité des erreurs ?
- Satisfaction : quel plaisir l'utilisateur tire lorsqu'il fait usage de l'interface ?

1. Il faut cependant noter que Qwant n'est pas un moteur de recherche à propre parler, mais un agrégateur de moteurs. Il retranscrit en fait les résultats des moteurs Bing (catégorie Web et Live), Kurrently (catégorie Social), Amazon (catégorie Shopping) et Wikipédia (catégorie Knowledge Graph) (source Numerama : http://www.numerama.com/magazine/25137-qwant-n-est-pas-un-moteur-de-recherche-mais-une-interface.html)

Enfin, la présentation des résultats d'une recherche dépend de deux paramètres : la pertinence système et la pertinence utilisateur. La pertinence système concerne directement l'appariement requête/documents et désigne le « score graduel traduisant la pertinence du document vis-à-vis de la requête » (Pinel-Sauvagnat, 2005). La pertinence utilisateur traduit l'intérêt de l'utilisateur pour une information donnée selon Mizzaro (1997). Cette notion est difficile à évaluer car c'est une notion « humaine » (Saracevic, 2008). Nous avons pourtant vu en 3.1.1 que l'évaluation des résultats par l'utilisateur fait partie intégrante du processus de recherche. Marchionini et White (2008) considèrent que cette évaluation peut donner lieu à une reformulation du besoin informationnel qui va se traduire par une seconde requête. Ils considèrent également que la pertinence est un des facteurs contextuels les plus importants en RI, interagissant avec d'autres facteurs comme les préférences de l'utilisateur.

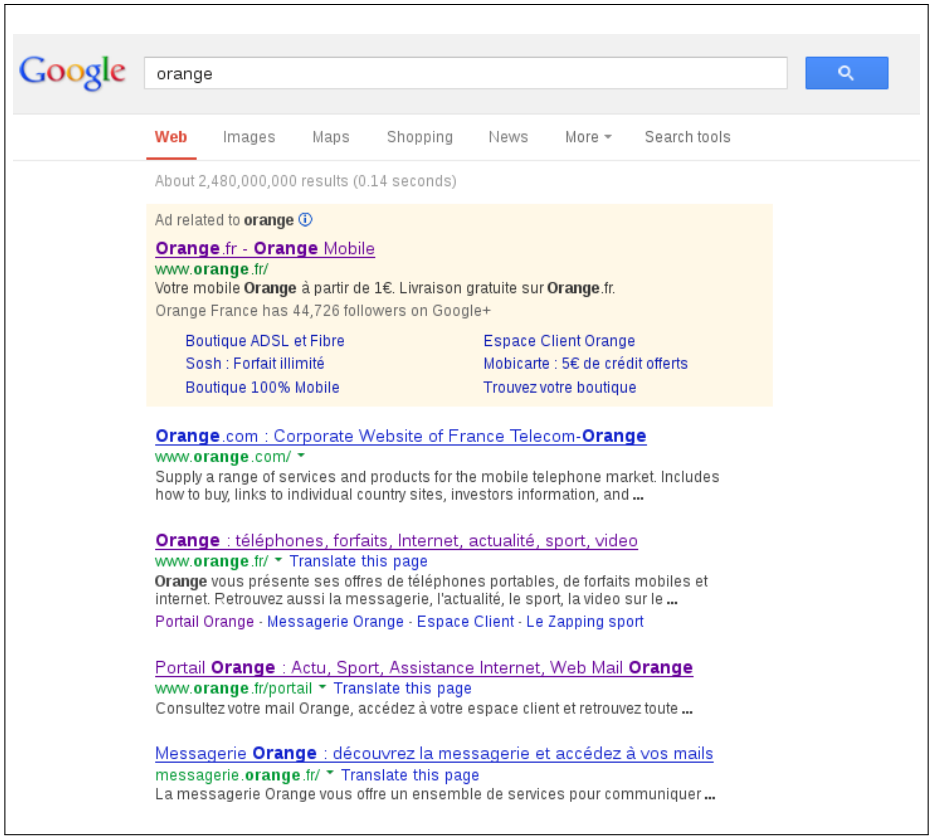


FIGURE 3.3 : Résultats sous forme de liste verticale à la requête *orange* (Google)

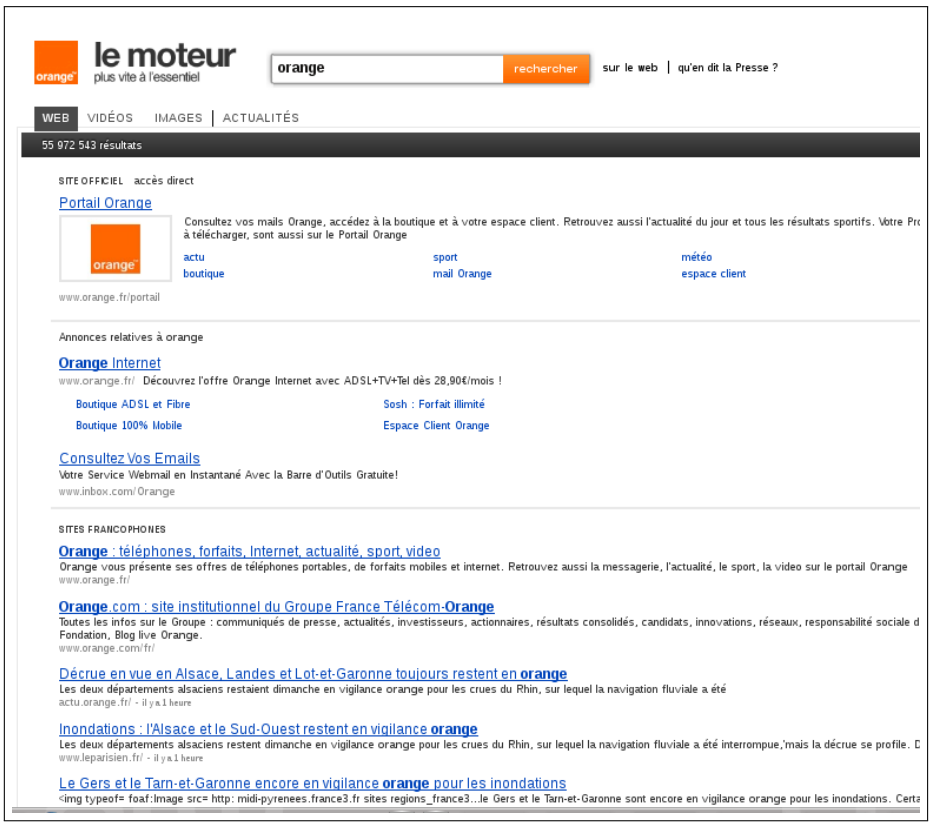


FIGURE 3.4 : Résultats sous forme de liste verticale à la requête *orange* (Orange)

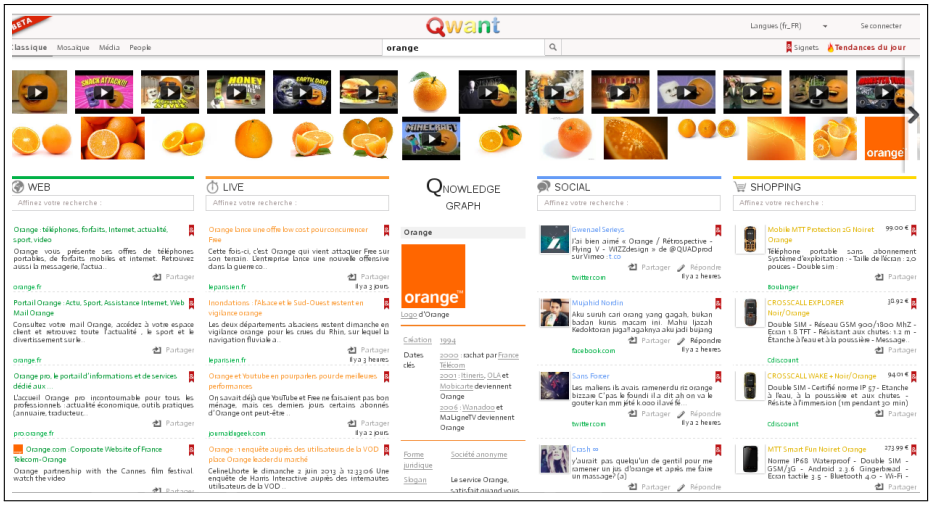


FIGURE 3.5 : Résultats classés par catégories à la requête *orange* (Qwant)

3.2 De la RI traditionnelle à la RI contextuelle

La RI contextuelle a été posée comme défi à long terme par Allan *et al.* (2003), l'objectif étant de combiner des technologies de recherche avec des connaissances sur les requêtes et le contexte utilisateur, afin de mieux répondre au besoin d'information des utilisateurs. La RI contextuelle place donc l'utilisateur au centre du système de recherche d'information.

3.2.1 La recherche d'information contextuelle

Mothe (2011) définit la recherche d'information contextuelle de la manière suivante :

> « La RI contextuelle vise à modéliser les différents aspects du contexte et leur variété pour les intégrer dans le processus de recherche. L'aspect contextuel fait référence à des connaissances implicites ou explicites concernant les intentions de l'utilisateur, l'environnement de l'utilisateur et le système lui-même. L'hypothèse est que rendre explicites certains élements du contexte de la RI pourrait améliorer les performances des systèmes de RI. »

Cette définition soulève deux questions : qu'est-ce que le contexte de la recherche d'information ? Comment intègre-t-on du contexte en RI ?

Le contexte de la recherche représente tous les éléments qui peuvent avoir un impact sur la recherche. Selon Chevalier (2011), le contexte de la recherche est composé de trois éléments : l'utilisateur, le système de recherche et la tâche de recherche. L'utilisateur met en jeu ses connaissances et ses compétences. Le système de recherche gère la manière dont les résultats sont retournés à l'utilisateur (quel nombre de documents, quelle efficacité, etc.). Enfin, la tâche de recherche représente le type de recherche effectuée : recherche d'une information connue, recherche d'une information quelconque, recherche exploratoire, recherche exhaustive (Morville et Rosenfeld, 1998). Le contexte en RI représente donc l'ensemble des éléments pouvant avoir un impact sur la RI. Le contexte est donc multidimensionnel et ses dimensions peuvent varier selon les auteurs.

Selon Mothe (2011), il existe deux types de tâches en RI contextuelle : la modélisation des différents aspects du contexte et l'intégration du contexte en RI. Nous développons la question de l'intégration en 3.2.3. La modélisation des différents aspects du contexte s'attache à rendre compte de manière analytique des différents acteurs et phénomènes qui constituent le contexte en RI.

Cette modélisation est présente dans le modèle analytique de Ingwersen et Järvelin (2005) visible dans la figure 3.6. Il rend compte des interactions qui existent entre les différents objets qui constituent le contexte d'un système de RI, objets qui sont fortement variables (Mothe, 2011).

Dans ce modèle, l'utilisateur (*Cognitive Actor*) est influencé par un certain nombre de facteurs dont le contexte social et culturel. L'interface est alors un vecteur d'interaction entre l'utilisateur et le système (*Engines logics Algorithm*) ainsi que les informations gérées par celui-ci. L'interface est également un élément qui doit pouvoir s'adapter aux conditions de recherche. Enfin, le système dépend de l'information qu'il gère (*Informations objects*) puisque la constitution du système va être adaptée en fonction de celles-ci.

La modélisation proposée par Ingwersen et Järvelin (2005) nous montre que les objets qui constituent le contexte sont le plus souvent interconnectés, variables et de nature différente.

Nous allons nous attacher à présenter ce qu'est le contexte en RI dans ce qui suit. Dans un premier temps, nous allons ainsi présenter la typologie de Tamine-Lechani (2008) qui synthétise les différentes dimensions du contexte décrites par les auteurs du domaine. Dans un deuxième temps, nous allons traiter de la RI contextuelle en elle-même, en nous focalisant sur l'intégration du contexte en RI à travers les travaux sur les requêtes qui restent les principales traces exploitables du contexte. Enfin, nous terminons avec le dispositif d'intégration du contexte qu'est la personnalisation en RI.

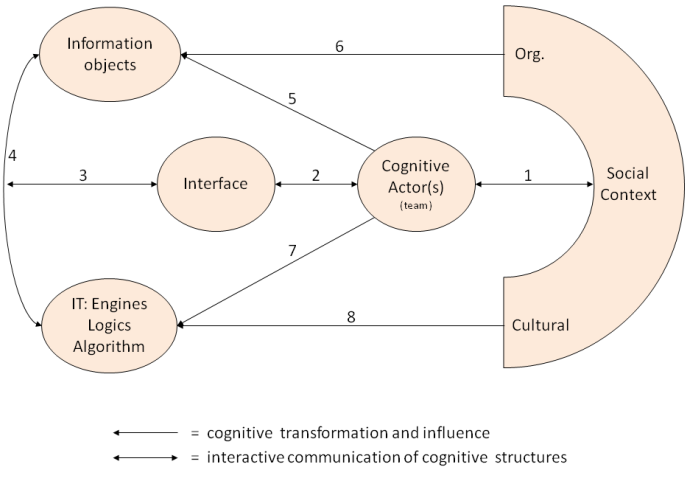


FIGURE 3.6 : Modèle analytique général de la recherche d'information par Ingwersen et Järvelin (2005)

3.2.2 Le contexte en RI

Selon Tamine-Lechani (2008), cinq dimensions ont un réel impact sur la qualité du processus de recherche et il est important d'adapter le processus de recherche à chacune d'elles. Ces dimensions sont discutées par les différents auteurs, il existe cependant un noyau commun : l'environnement et les dimensions humaines (Chevalier, 2011).

3.2.2.1 Les dimensions qui dépendent de l'environnement

Parmi les cinq dimensions du contexte décrites par Tamine-Lechani (2008), deux d'entre elles dépendent principalement de l'environnement qui entoure le processus de recherche : le contexte spatio-temporel et les moyens d'accès à l'information. Les moyens d'accès à l'information constituent l'environnement matériel qui permet à l'utilisateur d'utiliser l'application de recherche.

Le contexte spatio-temporel La localisation géographique et temporelle forme deux dimensions particulièrement importantes selon le type d'application considéré. En effet, les informations proposées par certains services de recherche ont une validité subordonnée au lieu et au temps instanciés. Les guides touristiques, les cartes interactives sont les services les plus concernés puisque leur fonction est justement d'aider l'utilisateur à se diriger (Kofod-Petersen et Aamodt, 2003) ; ou encore les systèmes de billeterie automatique où la variable temporelle est essentielle.

Les moyens d'accès à l'information Cette dimension se focalise sur les caractéristiques de l'outil physique permettant d'effectuer un accès direct à l'information (ordinateur, tablette numérique, téléphone portable, télévision, etc.). Les moyens d'accès à l'information affectent surtout la présentation des résultats, et la possibilité d'interagir avec l'outil. La prise en compte de cet aspect du contexte a généré de nouveaux modes d'accès comme la reconnaissance vocale de Google qui facilite l'utilisation du moteur à partir d'un téléphone portable. Cet aspect du contexte a un impact direct sur la tâche de recherche d'information.

3.2.2.2 Les dimensions humaines du contexte

Les dimensions humaines du contexte décrites par Tamine-Lechani (2008), sont au nombre de trois : l'utilisateur, l'intention de l'utilisateur (la tâche de

recherche) et la perception de l'utilisateur vis-à-vis de l'information proposée (le contexte de l'information).

Le contexte utilisateur Cette dimension peut être décomposée en deux aspects : le contexte personnel et le contexte social (Tamine-Lechani, 2008). Le *contexte personnel* contient beaucoup de variables. Il comprend des paramètres démographiques comme les facteurs de préférences personnelles tels que la langue ou le sexe de l'utilisateur (Hupfer et Detlor, 2006) mais aussi des paramètres d'ordre psychologique, qui peuvent affecter le jugement de pertinence de l'utilisateur (comme l'anxiété ou la frustration). Le contexte personnel recouvre également des questions de cognition comprenant l'expertise et les centres d'intérêt de l'utilisateur (Ingwersen et Järvelin, 2004).

Le *contexte social* met l'accent sur la communauté à laquelle l'utilisateur appartient : ses amis, ses voisins et ses collègues. Cette dimension a été développée par Ingwersen et Järvelin (2004). L'utilisateur est influencé dans sa recherche par son milieu social et culturel. L'influence sociale et culturelle a également un impact sur le type d'objet convoité par l'utilisateur.

Les variables mises en avant ne sont pas toutes facilement accessibles dans le but de les intégrer à un système. En effet, si la langue ou les centres d'intérêt de l'utilisateur peuvent être connus, il est beaucoup plus difficile de détecter et d'utiliser des informations sur l'état émotionnel de l'utilisateur.

La tâche de recherche La tâche de recherche porte sur l'intention de l'utilisateur et se matérialise par l'expression de son besoin d'information sous la forme d'une requête. Les expériences menées par Navarro-Prieto *et al.* (1999) montrent que la stratégie de recherche adoptée par l'utilisateur est fortement liée à la tâche de recherche. Si la tâche consiste à rechercher des informations non connues et non ciblées, l'utilisateur produit une recherche exploratoire. Celle-ci se traduit par l'utilisation de requêtes génériques. Par exemple, lorsqu'on donne comme consigne à un utilisateur de chercher des informations sur les maladies psychologiques, il va produire des requêtes comme *psychologie* ou *maladies* qu'il va par la suite affiner en y ajoutant des mots-clés. Au contraire, si la tâche consiste à effectuer une recherche ciblée comme par exemple chercher les offres d'emploi dans un domaine professionnel spécifique, l'utilisateur utilise des requêtes spécifiques. Potentiellement, l'aspect non ciblé d'une requête pourrait donc avoir un impact sur la production d'une requête de type générique et ambiguë.

Les requêtes issues de la tâche de recherche peuvent traduire différentes intentions : informationnelle, navigationnelle ou transactionnelle (Broder, 2002;

Jansen *et al.*, 2007). Lorsque la requête a comme but une transaction, l'utilisateur utilise le système de recherche pour passer une transaction, comme par exemple chercher une voiture d'occasion. Les requêtes informationnelles ont pour but de trouver et acquérir une information dans un ou plusieurs documents. Par exemple, on peut vouloir connaître les résultats des épreuves de saut en longueur aux Jeux Olympiques de 2012 et proposer une requête comme *saut en longueur JO2012*. Une requête navigationnelle consiste à utiliser le système de recherche dans le but d'accéder à un site particulier, par exemple proposer la requête *facebook* dans le but d'accéder au site du réseau social. Nous reviendrons sur ces aspects en 5.4.1.

L'identification de la valeur de la requête qui a été formulée (ciblée, informationnelle, navigationnelle, etc.) peut être un indice intéressant pour traiter l'ambiguïté d'une requête. En effet, deux requêtes identiques peuvent être la résultante de stratégies de recherche différentes, par conséquent les utilisateurs, auteurs des requêtes, n'attendront pas le même résultat. Ainsi la requête *facebook* peut être navigationnelle ou au contraire informationnelle. Dans le premier cas, l'utilisateur attendra l'adresse du site en résultat, dans le deuxième cas, il attendra des informations sur le site ou à propos de l'entreprise.

Le contexte de l'information Cette dernière dimension décrit le contexte direct de l'information, c'est-à-dire le document qui contient l'information et la représentation que peut en avoir l'utilisateur (Tamine-Lechani, 2008). Cette dimension a été mise en évidence par Ingwersen (1994), qui a proposé la notion de « polyreprésentation » de l'information. La polyreprésentation est une notion qui suppose que les informations sont classifiables selon un ensemble de critères tels que le genre de l'information, son auteur, sa structure, son style etc. Le document peut être caractérisé par sa mise en forme, sa couleur ou les métadonnées de sa structure.

La perception de l'information par les utilisateurs, élément constituant le contexte de l'information, a un impact sur la pertinence utilisateur (cf. 3.1.3). La perception de l'utilisateur vis à vis du contexte de l'information peut évoluer en fonction de la tâche de recherche qu'il accomplit (Tombros *et al.*, 2005). En effet, l'utilisateur, selon qu'il cherche à faire une recherche générale et exhaustive à propos d'un sujet ou à accomplir une tâche de décision en vue d'un achat, ou encore une tâche de compilation d'information en vue d'un voyage, n'utilisera pas les mêmes élements du contexte.

3.2.3 L'intégration du contexte en RI

L'intégration du contexte en RI peut intervenir à tous les niveaux décrits précédemment. Cependant, comme le souligne le collectif (Allan *et al.*, 2003), l'extraction et la représentation des connaissances concernant le contexte de la recherche d'information et les utilisateurs reste une tâche difficile. Les travaux se focalisent principalement sur les requêtes des utilisateurs car celles-ci sont plus accessibles mais comme le soulignent Allan *et al.* (2003) dans leur rapport, les requêtes ne sont pas les seules expressions des besoins d'informations. Les requêtes restent cependant les principales traces exploitables du contexte. Les requêtes et les historiques de recherche contiennent en effet de nombreuses informations concernant le contexte de la RI : informations sur la tâche de recherche, sur l'utilisateur ou encore sur le contexte de l'information [2]. C'est pourquoi nous proposons ici de détailler deux types de requêtes intéressantes du point de vue de l'intégration du contexte en RI : les requêtes répétées (cf. 3.2.3.1) et les requêtes reformulées (cf. 3.2.3.2).

3.2.3.1 Les requêtes populaires ou répétées

Le cas des requêtes fréquentes est intéressant pour les systèmes de RI, car elles peuvent être traitées par des méthodes spécifiques de filtrage de l'information ou des principes de recommandations (Mothe, 2011). Selon Zhang et Lu (2009), il existe deux types de récurrence :
– les récurrences collectives qui sont formulées par différents utilisateurs ;
– les récurrences individuelles produites par un même utilisateur.
Cette distinction permet de considérer deux types de phénomènes. Dans le premier cas, on peut parler de requêtes « populaires » alors que dans le second cas, on peut parler de requêtes « répétées » où le caractère de nouveauté semble déterminant. Le deuxième cas est typique d'une requête navigationnelle (Teevan *et al.*, 2006). Cependant, ce type de distinction ne peut se faire sans l'utilisation de logs de requêtes non anonymisées, comprenant l'ensemble des traces produites par chaque utilisateur.

Par ailleurs, plusieurs travaux montrent que des requêtes sont fortement répétées. Cela touche une proprotion de 21,9% pour Tyler et Teevan (2010) (requêtes du moteur Live search) et jusqu'à 50% pour Sanderson et Dumais (2007). Toutefois, la notion de « requêtes répétées » varie selon les auteurs. Certains comme Teevan *et al.* (2006) considérent qu'une requête est répétée lorsqu'un utilisateur U_1 produit une requête X et clique sur un lien Y, puis

2. Voir 5.1.4 pour un exemple d'historique de recherche

qu'un utilisateur U_2 produit une requête identique à U_1 et clique sur le même lien. Mais les utilisateurs peuvent également cliquer sur des liens différents ce qui arrive dans 38% des cas dans le corpus de Teevan *et al.* (2006). En cherchant à identifier les requêtes répétées par un même utilisateur, ces travaux essayent de distinguer un type de requêtes particulières que sont les requêtes navigationnelles (*cf.* 3.2.2.2 et 5.4.1).

Sanderson et Dumais (2007) trouvent que 50% des requêtes étudiées sont répétées, et ce parce qu'ils ont une définition de la répétition de requêtes différente de celle de Teevan *et al.* (2006). En effet, ces auteurs étudient des requêtes anonymisées ce qui rend impossible de savoir avec exactitude si deux requêtes identiques sont produites par le même utilisateur. Ils ont donc cherché à reproduire des sessions utilisateurs, en considérant les requêtes répétées dans une fenêtre temporelle de 30 minutes.

Les répétitions sont donc de différents types, il est important de les distinguer pour apporter des précisions essentielles pour améliorer le système de RI comme, par exemple, mettre en place une personnalisation des résultats dans le cas de requêtes navigationnelles fréquentes.

3.2.3.2 Les requêtes reformulées

Le phénomène de reformulation des requêtes est très important en RI. Ainsi, Spink *et al.* (2002a) signalent que 52% des utilisateurs reformulent leurs requêtes dans le corpus Excite 97 et 45% dans le corpus Excite 2001. Ces reformulations peuvent être de précieux indices sur le besoin d'information de l'utilisateur. En effet, les reformulations peuvent être de différentes formes. Les requêtes peuvent être étendues (en ajoutant un ou deux mots) ou réduites (en enlevant un ou deux mots). Les requêtes reformulées par extension représentent 20% des requêtes reformulées selon Jansen *et al.* (2000). Les requêtes reformulées par diminution du nombre de terme correspondent à 25% des requêtes reformulées. Et on relève que 35% des requêtes sont reformulées sans changement de longueur.

Les mouvements d'extension ou de réduction des requêtes se traduisent par quatre types de mécanismes (Adam *et al.*, 2013; Huang et Efthimiadis, 2009) :
- généralisation : ce mécanisme va consister à supprimer un ou plusieurs mots comme dans *dates concert franz ferdinand* $\rightarrow$ *franz ferdinand*, on peut également remplacer certains mots par un hyperonyme (*évènements franz ferdinand*) ;
- spécification : cela se traduit par un mouvement inverse à la généralisation : *pop rock* $\rightarrow$ *groupes anglais pop rock* ;

– reformulation : il s'agit à proprement parler d'une paraphrase. Elle va consister à remplacer à des mots de la requête par un synonyme comme dans *nouveau album franz ferdinand → nouveau disque franz ferdinand* ;
– mouvement parallèle : ce mouvement crée une modification importante de la requête créant une alternative comme le signalent Adam *et al.* (2013), par le remplacement par exemple d'un produit ou d'une marque par un autre (*nouveau album franz ferdinand → nouvel album phoenix*).

Toutefois, la reformulation de requêtes peut concerner des phénomènes plus difficilement quantifiables et repérables automatiquement. Par exemple Rieh et Xie (2006) signalent que 80% des reformulations de leur échantillon d'étude (issu du moteur Excite en 2000) concernent le contenu. Rieh et Xie (2006) quantifient les différents types de reformulation : généralisation (15,8%), spécialisation (29,1%), remplacement par un synonyme (4,9%) ou mouvements parallèles (51,4%). Mais les critères déterminant ce qu'est un mouvement parallèle restent peu clairs dans cette étude.

Selon Kumaran et Allan (2008), une requête relaxée va donner lieu à la création de sous-requêtes peu efficaces en terme de recherche, contrairement à une requête courte qui va être étendue. Ce type de connaissance sur les requêtes permet de réaliser des expansions de requêtes automatiques de meilleure qualité.

De même, le repérage des reformulations d'une requête permet d'étendre le contexte de la requête, en précisant par le biais des autres requêtes reformulées le besoin d'information. Ainsi d'après Shen et Zhai (2003), ce procédé peut avoir un intérêt pour désambiguïser certains mots des requêtes, ce que nous avons également vu en 2.2.2.3. Shen et Zhai (2003) montrent que l'utilisation de l'historique de recherche des utilisateurs permet d'améliorer les performances (en particulier la précision) lorsqu'elle est combinée aux clics des utilisateurs.

3.2.3.3 La personnalisation

Les moteurs de recherche commerciaux produisent, pour la plupart, des résultats identiques pour tous les utilisateurs. En effet, les moteurs de recherche sont conçus pour satisfaire les besoins du groupe plutôt que les besoins individuels (Teevan *et al.*, 2005). Cependant, comme le signale (Hearst, 2009), la personnalisation des résultats de recherche n'est qu'une forme de personnalisation parmi d'autres en RI. Elle propose d'ailleurs une classification des différentes formes de personnalisation, en distinguant ce qui relève de la per-

sonnalisation au niveau individuel et au niveau collectif. Elle distingue également deux types de personnalisation :

– les personnalisations explicites qui demandent à l'utilisateur d'agir et préciser leurs choix (création d'alertes sur des sujets précis) ;
– les personnalisations implicites qui sont faites automatiquement (analyse de l'historique de recherche ou réordonnancement des résultats en fonction des choix des utilisateurs).

La personnalisation explicite va consister en un paramétrage ou une « customisation » du sytème, où l'utilisateur configure le système via des options ou des préférences intégrées par les concepteurs du système (Chevalier, 2011), alors que la personnalisation implicite va consister en une adaptation dynamique du système. Le système collecte les informations disponibles, les analyse puis les synthétise. Le système s'adapte en appliquant une décision ou un traitement spécifique en fonction des informations collectées. Contrairement au paramétrage où les préférences de l'utilisateurs sont intégrées au système, la personnalisation implicite d'un système demande la création d'un espace appelé « profil utilisateur ». Ce profil stocke les informations concernant l'utilisateur. Bouzeghoub et Kostadinov (2005) listent les informations qu'il peut contenir :

– Données personnelles ;
– Historique des interactions de l'utilisateur ;
– Centres d'intérêt ;
– Qualité attendue ;
– Sécurité.

C'est dans ce cadre que se déroulent les travaux exposés en 2.2.2.4 qui rapprochent la question de l'ambiguïté et de la personnalisation (Dou *et al.*, 2007; Teevan *et al.*, 2008). Ces travaux utilisent des éléments contextuels qui relèvent d'une personnalisation à un niveau collectif et implicite, en l'occurrence les historiques de recherche des utilisateurs qui contiennent à la fois les requêtes et les documents cliqués par les utilisateurs. Nous développons ces aspects dans les chapitres 4 et 5. Toutefois, Hearst (2009) modère l'impact de ces méthodes qui ne peuvent être appliquées sur les requêtes qui n'ont jamais été soumises auparavant. En effet, la personnalisation des résultats va être pertinente pour les requêtes répétées (environ 50% des cas *cf.* 3.2.3.1). Et Hearst (2009) précise également que les méthodes analysant les clics à un niveau collectif ne sont pas en mesure de donner des bons résultats dès lors qu'une requête peut avoir plus d'une interprétation. L'ambiguïté d'une requête demande un traitement personnalisé, à un niveau individuel, nécessitant alors l'utilisation d'indices contextuels au plus près de l'utilisateur et de son interaction avec le système.

3.3 Conclusion

L'évolution de la RI correspond à la fois à une adaptation à un nouveau contexte technologique mais aussi à des changements d'attente de la part des utilisateurs, et à l'apparition de nouvelles perspectives de recherche. De plus, le Web ne cesse de grandir, les systèmes de recherche d'information se multiplient et se spécialisent de plus en plus, chaque utilisateur ayant des besoins informationnels différents.

L'étude du contexte permet de mieux connaitre les pratiques des utilisateurs comme, par exemple, la forte proportion de requêtes navigationnelles répétées. Ces études rendent également possible l'intégration de dispositifs qui adaptent le système à l'utilisateur. L'intégration du contexte en RI permet la mise au jour de phénomènes « génériques » comme la reformulation de requêtes ou la répétion de certaines requêtes. Ce sont autant d'indices précieux pour détecter des phénomènes liés aux requêtes des utilisateurs comme l'ambiguïté. Si chaque système de RI a une structure contextuelle similaire, les propriétés de ce contexte vont varier selon le type d'application de RI ou le public visé, ce qui ne rend pas son exploitation facile et systématique. La prise en compte du contexte est donc essentielle dans notre approche, en mettant l'accent sur les données issues des systèmes de RI et sur les productions des utilisateurs, au même titre que les processus techniques de la RI.

Deuxième partie

Un moteur de recherche spécialisé : 2424actu

Chapitre 4

Moteurs de recherche spécialisés : le cas de l'accès à l'actualité

Ce chapitre traite particulièrement du contexte de la thèse, qui a été réalisée dans un cadre industriel. Notre travail porte sur le moteur de recherche intégré au site 2424actu.fr, développé à Orange Labs. L'objectif de ce chapitre est de comprendre l'importance et la particularité de ce contexte industriel. En effet, chaque moteur de recherche a ses propres caractéristiques. La recherche d'information ne se pratique pas de la même manière selon les utilisateurs et l'information visés. Afin de comprendre et d'expliciter les specificités de ce cadre de travail, nous proposons d'aborder tout d'abord la question des moteurs de recherche spécialisés. Puis, dans un deuxième temps, nous nous focalisons sur le domaine auquel donne accès le moteur de recherche de 2424actu, l'actualité. Enfin, nous terminons cette caractérisation du moteur par une modélisation du contexte qui l'entoure, ce qui permet de mettre en évidence ses particularités techniques.

4.1 Caractéristiques d'un moteur spécialisé

L'accès à l'actualité sur internet est rendu possible aujourd'hui par l'utilisation de moteurs de recherche dédiés à l'actualité. Chaque moteur a ses spécificités ; les comprendre permet d'expliciter certains comportements des utilisateurs d'un moteur. Nous allons pour cela recourir à une typologie des moteurs de recherche, qui nous permettra alors de situer le moteur sur lequel nous avons travaillé par rapport aux autres technologies similaires.

Différentes typologies sont envisageables. En effet, on peut organiser les différents types de moteurs selon la façon dont sont organisés les résultats pro-

posés, en distinguant, par exemple, les moteurs dont les résultats sont fournis sous forme de listes ordonnées comme Google et les moteurs qui proposent des résultats organisés par catégories (cf. section 3.1.3) ou sous forme cartographique comme Cluzz [1]. On peut également typer les moteurs selon le degré d'interactivité qu'ils ménagent à l'utilisateur, à savoir s'ils autorisent la participation des utilisateurs, et de quelle manière. Cependant le critère le plus utilisé depuis l'apparition des premiers moteurs de recherche sur le Web au début des années 1990 est le caractère générique ou spécifique d'un moteur. Cette typologie, basée sur les ressources collectées, distingue deux types de moteurs :

– les moteurs généralistes ou horizontaux (*general-purpose engine*) ;
– les moteurs spécialisés ou verticaux (*domain-specific engine*).

Les moteurs généralistes parcourent le Web et proposent aux utilisateurs un accès à des milliards de documents. C'est le cas de Google, Yahoo ! ou Bing. Les moteurs spécialisés donnent quant à eux accès à des ressources spécifiques. L'évolution du Web, au cours des vingt dernières années, a largement confirmé cette dichotomie initiale. En effet, le premier moteur généraliste WWWW (McBryan, 1994) indexait 110 000 pages web en 1994 et recevait le nombre reccord de 1500 requêtes par semaine. En 1997, Altavista recevait 20 millions de requêtes par jour (Brin et Page, 1998). L'augmentation fulgurante à la fois de la couverture du Web et du nombre d'utilisateurs a creusé les différences entre ces deux types de moteurs.

Les moteurs spécialisés sont ainsi des modules spécialisés à l'intérieur des moteurs généralistes. Google a développé un grand nombre d'applications de ce type. L'URFIST de Rennes (Unité Régionale de Formation à l'Information Scientifique et Technique) [2] propose une typologie détaillée des moteurs spécialisés :

– selon les ressources internet : il peut s'agir de listes de diffusion comme par exemple Tile.net, ou de forums (1001Forums), ou de blogs comme dans le cas de Wikio. Certains moteurs peuvent également offrir la possibilité d'accéder à des contenus provenant des réseaux sociaux ou des plates-formes de microblogging (comme le moteur de Twitter ou Topsy).
– selon la nature du média : certains moteurs sont spécialisés dans la recherche d'images (comme par exemple Pixolution), de vidéos comme le moteur de Dailymotion ou de fichiers au format .pdf (PdfGeni).
– selon la nature du contenu : par exemple, certains moteurs permettent de chercher des actualités, des documents scientifiques, des personnes ou encore des contenus géolocalisés. Ainsi les moteurs dédiés à la recherche d'ar-

1. Les urls des moteurs cités sont disponibles en Annexe A.1 (page 207).
2. http://www.sites.univ-rennes2.fr/urfist/node/320

ticles scientifiques sont nombreux (Scirius, CiteseerX) et ils peuvent être spécialisés dans un domaine particulier. Les moteurs EEM ou GoogleNews permettent d'accéder spécifiquement à des actualités. C'est le cas du moteur de recherche 2424actu, alimenté par un agrégateur d'actualité.

4.2 Le cas de l'accès à l'actualité en ligne : l'agrégateur 2424actu

Nos travaux se sont basés sur la plate-forme 2424actu.fr, déployée du 1er octobre 2009 au 30 septembre 2011. Cette plate-forme a vécu seulement deux ans sous cette forme puis a été transférée sous le portail Orange [3]. Nous avons dû composer avec de nombreux changements techniques mais également de nombreux changements humains, ce qui a rendu difficile la collecte d'informations sur l'application en elle-même. La plate-forme 2424actu n'a cessé d'évoluer durant ces 2 ans : arrivée et départ de nouveaux fournisseurs, ajout de nouveaux traitements, déploiement technique sur d'autres supports.

Cette plate-forme est un agrégateur en ligne qui comprend également un moteur de recherche. Ce moteur donne exclusivement accès à des informations relevant du domaine de l'actualité. Ce moteur est spécialisé dans l'accès à l'actualité francophone. En ce sens, il correspond à la définition d'un moteur spécialisé vu en 4.1. Du point de vue de la typologie présentée précédemment, le moteur 2424 se distingue sur deux aspects :

- par la nature du contenu : il est spécialisé dans l'accès à un domaine : l'actualité ;
- par la nature du média car il est spécialisé en particulier dans la mise à disposition des vidéos d'actualité.

Dans ce qui suit, nous présentons tout d'abord ce que sont les agrégateurs d'actualité, en particulier leur fonctionnement vis-à-vis de la presse. Dans un second temps, nous nous intéressons à l'agrégateur 2424actu en lui même, en décrivant les modalités d'accès à l'actualité qu'il propose.

4.2.1 Les agrégateurs d'actualité

Le site 2424actu.fr est un *agrégateur d'actualité*. Selon Boure et Symrnaios (2006), un agrégateur d'actualité a une « activité d'infomédiation » dans le domaine de l'actualité. L'infomédiation désigne la « fonction consistant à relier des besoins ciblés et des ressources pertinentes au sein de volumes de données considérables et hétérogènes » (ibid, 2006). Rebillard et Smyrnaios (2010) précisent

3. http://www.actu.orange.fr/revue/actualite-du-jour/

que l'infomédiation ne se limite pas à une activité de recherche de données, elle apporte un service de mise à disposition de contenus et de liens, issus d'une sélection éditorialisée.

Les agrégateurs se caractérisent par une automatisation complète du processus, depuis la récupération des données d'actualité jusqu'à la présentation organisée de ces données à l'utilisateur. Ils sont surtout l'apanage des fournisseurs de services internet. En effet, les agrégateurs retravaillent des données récupérées en amont par des robots de recherche d'information qui parcourent inlassablement les sites proposant de l'information journalistique.

Les contenus proposés par les agrégateurs s'organisent en général en une Une et diverses rubriques spécialisées. Les agrégateurs se distinguent des portails par un agencement automatique de contenus, qui se traduit par l'absence d'une intervention éditoriale. Les sites portails sont des lieux qui concentrent de multiples flux de données et d'utilisateurs, avec une vérification humaine (Rebillard, 2006).

Les agrégateurs s'alimentent à plusieurs sources d'information. Au coeur de ce dispositif figurent les agences de presse, en l'occurence l'Agence France Presse (AFP) pour les agrégateurs francophones. Les agences de presse se situent toutes en amont de la production de l'information journalistique. A l'origine, elles ne produisaient des informations qu'à destination des professionnels du domaine. Mais l'expansion du web a permis à ces agences une diffusion de leurs dépêches à une plus grande échelle. Ces agences ont également la particularité de fournir des contenus dépourvus de ligne éditoriale ce qui donne une impression de neutralité (Rebillard, 2006).

Les contenus proviennent également de sites d'actualité généralistes et de sites appelés *pure player*. Les sites généralistes génèrent un grand nombre d'informations sur des temporalités rapides. Ils traitent en simultané un grand nombre de domaines. Ils émanent de médias déjà existants, que ce soit la télévision (France Télévision), la radio (RFI) ou bien la presse écrite papier, les quotidiens (LeMonde.fr, Liberation.fr, Lefigaro.fr, etc.) ou les hebdomadaires (Nouvelobs.fr). Les sites sont alimentés tout au long de la journée, soit par des éditions successives, soit en continu en suivant le flux des évènements. Les *pure player* désignent les entreprises qui exercent sur un seul domaine d'activité et par extension ce sont les sites web d'information sans édition papier. Ils se caractérisent par une grande souplesse quant à la longueur, au style et à l'écriture de chaque article publié (par exemple Rue89). Ils peuvent aussi intégrer du son, des images ou même des vidéos, le texte devenant alors secondaire. Chaque article est signé, que l'auteur soit un journaliste, un pigiste ou même un contributeur externe.

Enfin, le rythme de publication est différent par rapport à un site de médias d'actualité, moins d'articles sont publiés quotidiennement. Les articles publiés sont maintenus durant plusieurs jours (Charon et Floch, 2011). Les sources alimentant l'agrégateur sont donc très différentes, certaines ayant un contenu éditorial, d'autres produisant des articles courts, d'autres des articles d'enquêtes de plusieurs pages.

4.2.2 L'agrégateur 2424actu

L'agrégateur 2424actu se conforme à ces principes généraux. Mais il se démarque des autres agrégateurs (en particulier Google News [4]) par ses rapports avec les fournisseurs d'information journalistique. Tous les contenus du site proviennent de partenaires qui sont sélectionnés en amont et non pas de contenus récoltés sur le web.

Une autre particularité de l'agrégateur 2424actu est de proposer la vidéo comme point d'entrée dans l'actualité. Ce support d'information est en effet proposé en second plan dans les autres agrégateurs du marché, qui préférent la presse écrite.

L'application 2424actu comprend un agrégateur et un moteur de recherche. L'agrégateur en amont récupère les flux des fournisseurs de contenus d'actualité [5], comme par exemple Lemonde.fr ou l'AFP vidéo. Il agrège ensuite ces contenus, en opérant des regroupements de documents en fonction de leur contenu. Par exemple, des documents relatant des évènements à propos du tremblement de terre au Japon seront rassemblés automatiquement au sein d'un même groupe de documents. Le moteur de recherche permet aux utilisateurs d'avoir accès aux contenus agrégés via la soumission de requêtes.

La classification automatique est renouvelée toutes les vingt minutes. Le regroupement est complété par un recours à des catégories thématiques issues des étiquettes attribuées par l'AFP à chaque document. Ces étiquettes sont décrites dans la section 5.1.1.

La double organisation (clusters vidéo et thématiques) des documents porteurs d'informations d'actualité donne à l'utilisateur un accès à l'information par « plusieurs portes ». Au total, l'agrégateur 2424actu propose trois accès différents à l'information (4.1) :

– Accès direct via la « dalle » de vignettes d'images à la Une : cet accès permet à l'utilisateur d'avoir une *vue* sur l'actualité. La Une présente à l'utilisateur

4. http://news.google.fr/
5. Pour des raisons de confidentialité, la liste complète des sources ne peut être communiquée.

tous les *clusters* de documents considérés comme importants à un instant *t*. L'importance est déterminée par un ensemble de facteurs contextuels, le facteur principal étant la présence d'un document dans un format vidéo.

– Accès indirect via les différentes catégories thématiques de l'actualité : ces rubriques thématiques issues des étiquettes AFP donnent un accès à une *vue* spécialisée sur l'actualité.
– Barre de recherche via le moteur de recherche spécialisé : située en haut à droite de la figure 4.1, la barre de recherche permet à l'utilisateur de rechercher des informations dans la base de 2424actu.

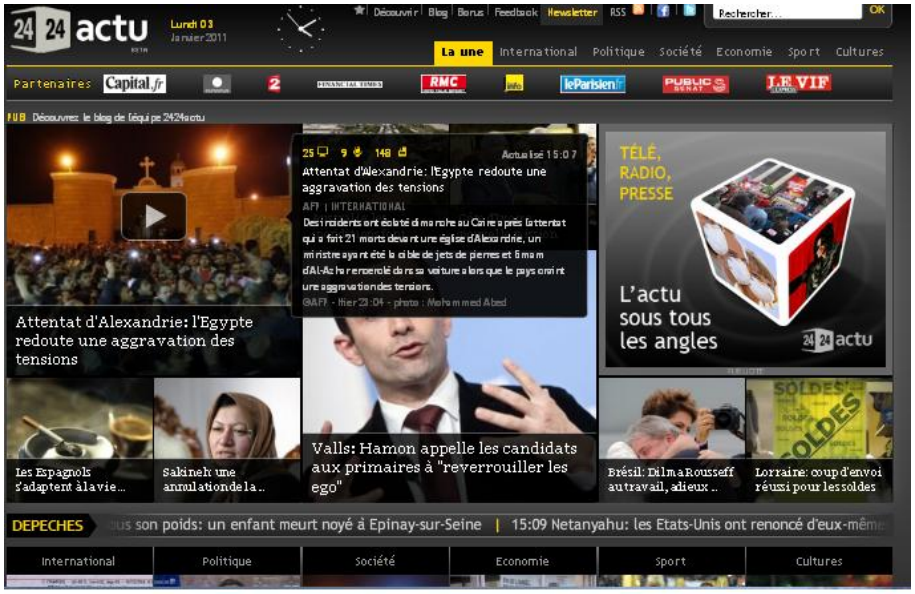


FIGURE 4.1 : 2424actu.fr (3/01/2010)

4.3 Modélisation de l'accès contextuel à l'actualité

Le moteur 2424actu est donc un moteur spécialisé du fait de la nature du contenu qu'il met à disposition des utilisateurs. C'est par ailleurs un moteur qui donne accès à l'actualité sous différentes formes. Nous reprenons ici les cinq dimensions principales du contexte en RI, décrites dans la typologie de Tamine-Lechani (2008) et exposées dans la section 3.2.2 du chapitre 3. Ces dimensions vont nous permettre de modéliser le contexte du moteur de recherche de 2424actu. Cette modélisation est un préalable au recueil des données présentées dans le chapitre 5. Pour décrire ces dimensions, nous nous appuyons sur les éléments connus du contexte du moteur 2424.

4.3.1 Les moyens d'accès à l'information

Les moyens d'accès à l'information construisent les élements de contexte dans lesquels s'intègre la recherche d'information. Au cours de ses deux ans d'existence, l'application 2424actu a été accessible sous différents supports. Ces supports conditionnent l'accès à l'information. En effet, les utilisateurs ont pu utiliser l'application et le moteur de recherche sur trois supports : ordinateur, tablette, téléphone (cf. figure 4.2). Toutefois, il n'est pas possible de savoir a posteriori quel support est utilisé par un utilisateur.

FIGURE 4.2 : Moyens d'accès à l'information sur 2424actu

4.3.2 Le contexte spatio-temporel de l'application

Les dimensions de localisation géographique et temporelle sont particulièrement importantes dans la modélisation du contexte de 2424actu. En effet, comme on peut le voir sur la figure 4.3, ces deux dimensions sont particulièrement développées. La dimension temporelle est certainement la plus marquante pour 2424actu. En effet, l'application est alimentée par l'actualité en temps réel. Les documents sont affichés en fonction de leur date de production. La question de récence est au cœur de l'application.

La dimension spatiale est aussi très présente. Elle se matérialise à travers le choix des sources d'information. 2424actu agrège de nombreux médias francophones qui ne sont pas français (Belgique, Canada, etc.). Des sources anglophones ont également été intégrées mais elles restent minoritaires. Enfin, il existe une part du contexte non accessible à l'utilisateur, mais présente. En effet, les informations proposées par l'application 2424 sont labellisées par des catégories de l'AFP (cf. 4.2.2), et les catégories comportent des informations spatiales. En effet, la catégorie INTERNATIONAL situe les informations « hors de France », en opposition avec la catégorie SOCIÉTÉ, qui situe les évènements en France.

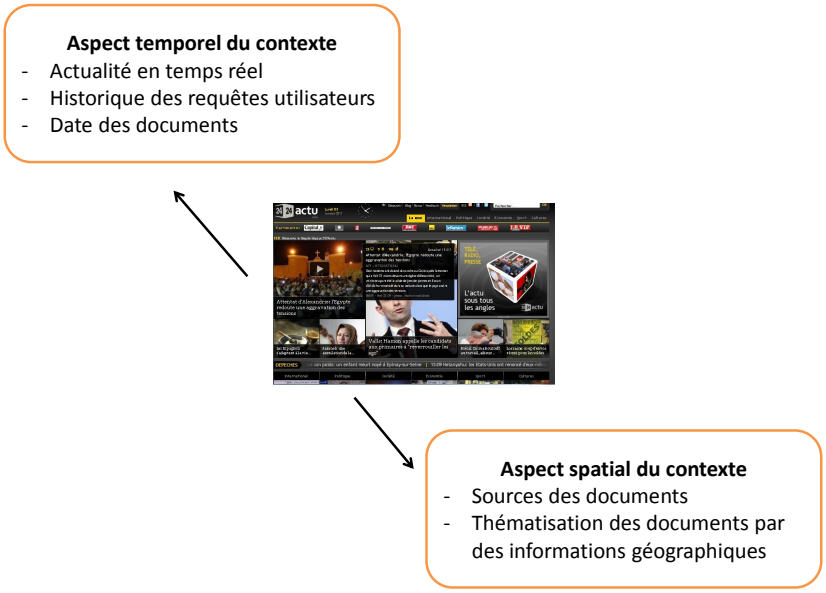


FIGURE 4.3 : Contexte spatio-temporel dans 2424actu

4.3.3 Le contexte utilisateur

Le contexte utilisateur est composé de deux dimensions : le contexte personnel et le contexte social. Ces deux dimensions présentes dans le cadre de l'application 2424actu sont représentées dans la figure 4.4. Le contexte personnel est très difficile à modéliser du fait de l'absence d'informations à ce sujet. On ne dispose pas d'information à propos des préférences des utilisateurs de 2424actu.

Le contexte social met l'accent sur la communauté d'appartenance de l'utilisateur. Orange a effectué une enquête marketing en amont du déploiement de l'application 2424. Cette enquête informe sur le type de public intéressé par ce type d'application et non pas sur les utilisateurs effectifs de l'application. Suite à cette enquête, on sait que le public intéressé est majoritairement composé des cadres supérieurs (noté CSP+ dans le schéma 4.4) qui ont une certaine attente en terme d'information en temps réel.

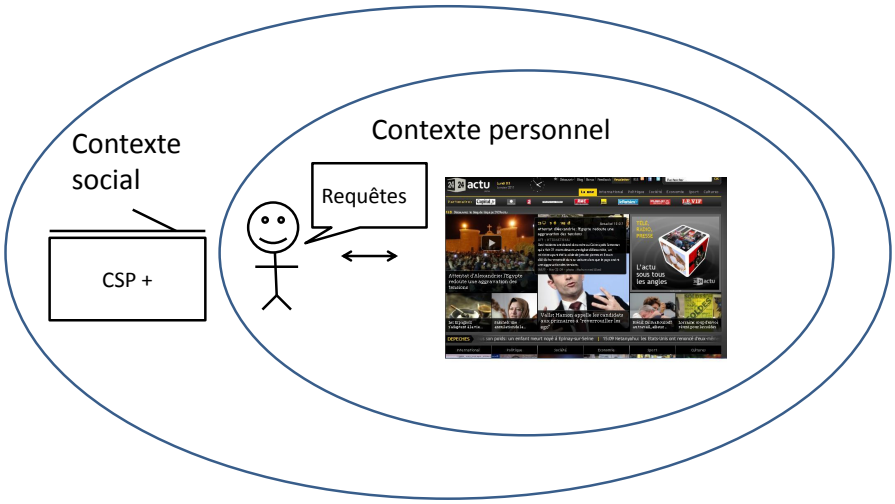


FIGURE 4.4 : Contexte utilisateur dans 2424actu

4.3.4 La tâche de recherche

Le contexte de la tâche de recherche met en évidence l'intention de l'utilisateur, cette intention est présente dans l'expression de sa requête. Le type de requêtes formulées informe donc sur le contexte de la tâche de recherche. Le site 2424actu permet de formuler trois types de requêtes visibles sur la figure 4.5 :

– rechercher une actualité ou une information ;
– rechercher une source ou un média particulier ;
– rechercher un élément propre au site (navigation intra-site).

Les possibilités de recherche du moteur 2424actu amènent donc l'utilisateur à formuler des requêtes principalement dans un but informationnel, et plus

rarement dans un but de navigation.

FIGURE 4.5 : Contexte de la tâche de recherche dans 2424actu

4.3.5 Le contexte de l'information

En modélisant cette dimension du contexte, on s'intéresse au contexte direct de l'information : sa forme et sa source. La figure 4.6 schématise ce contexte direct de l'information dans le cadre de 2424actu.

La forme de l'information dans 2424actu se matérialise sous différents formats : texte, vidéo et audio. Des métadonnées sont également présentes à travers des catégories thématiques (cf. section 5.1.1), et un classement des documents en fonction de leur récence.

Enfin, la source a une place prépondérante. Les sources journalistiques sont mises en avant dans l'interface de l'application, grâce à un bandeau de type publicitaire. Chaque source a une image a priori pour les utilisateurs, en termes de notoriété, de fiabilité, d'objectivité, qui conditionne le degré de confiance que l'utilisateur accorde à cette source.

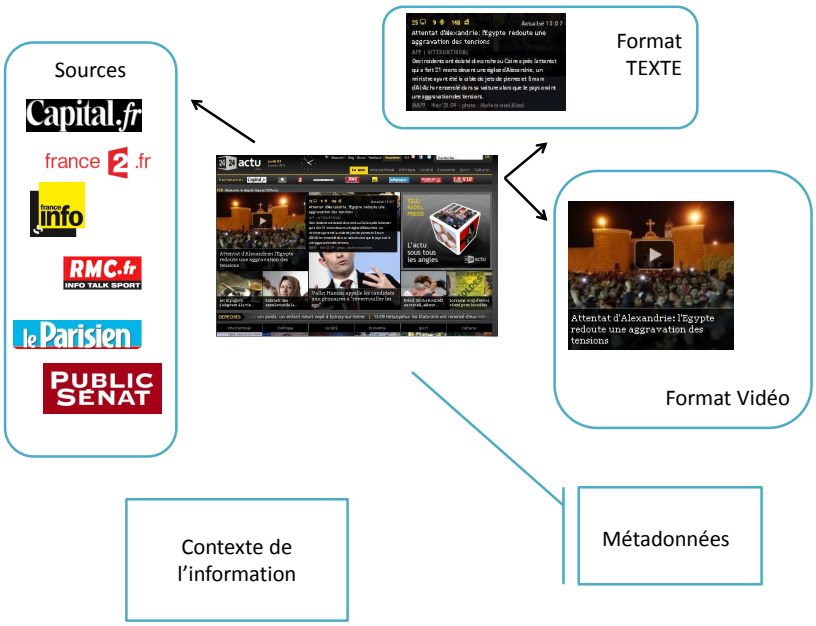


FIGURE 4.6 : Contexte de l'information dans 2424actu

4.4 Conclusion

Notre contexte de travail comprend donc à la fois un agrégateur d'actualité et un moteur spécialisé. L'application 2424actu regroupe tous ces composants. Elle se décline sous différents supports (web, applications pour tablettes et smartphones). Le moteur 2424 donne accès à un certain type de ressources, en l'occurrence de l'actualité. Nous avons vu la grande diversité des informations d'actualité que 2424 met à disposition. En effet, les producteurs de ce type de contenu sont variés et la particularité d'une application comme 2424actu est de mettre ensemble des contenus éditoriaux et non éditoriaux, et de les présenter de la même manière aux utilisateurs. L'inconvénient d'une telle démarche est de créer un environnement hétérogène. Cela va forcément influencer et contraindre la collecte des données nécessaires à notre travail comme on va le voir dans le chapitre suivant.

En nous appuyant sur le chapitre 3, nous avons modélisé le contexte de notre application 2424. Cette modélisation révèle de nombreuses traces contextuelles disponibles et accessibles. Les principales traces de ce contexte sont les requêtes des utilisateurs disponibles grâce aux historiques de recherche. Toutefois ce ne sont pas les seules traces exploitables, nous verrons dans le chapitre

suivant comment récolter ces traces dans un premier temps, puis comment les caractériser, étape préalable à l'utilisation de ces traces comme des données contextuelles.

Chapitre 5

Données et contraintes applicatives

Le but de ce chapitre est de rendre compte des différentes étapes qui ont permis de rendre exploitables les traces contextuelles et les données issues de l'application 2424. Ces étapes sont nécessaires pour pouvoir observer les requêtes qui sont adressées au système que nous étudions, afin de déterminer quelle forme peut prendre l'ambiguïté de requêtes dans un contexte applicatif. Le contexte industriel a orienté le choix des données vers une application existante. Cette démarche permet d'accéder à un contexte réel intéressant et très riche. Pour mener à bien cet objectif, nous avons dû réaliser un important travail de collecte et de caractérisation. La caractérisation est une étape de qualification des données, nous permettant d'identifier leurs spécificités.

Nous explicitons dans ce chapitre les contraintes applicatives qui ont fortement influencé notre travail et les collectes des données. Nous présentons ensuite le corpus constitué et ses caractéristiques. Cette présentation se déroule en deux temps : dans un premier temps, nous décrivons les documents constitutifs du corpus. Puis dans un second temps, les requêtes des utilisateurs sur des périodes temporelles similaires. Enfin, nous proposons une première caractérisation des requêtes utilisateurs de 2424actu à travers la mise en lumière des pratiques de recherche. Cette caractérisation est complétée par une analyse du contenu linguistique des requêtes. Nous terminons par une analyse des profils temporels des requêtes. Ces analyses permettent de mieux connaitre les requêtes issues de 2424actu.

5.1 Les données de départ

Pour présenter les données sur lesquelles se basent nos expérimentations, nous allons débuter par une description des passeurs de données que sont le

moteur 2424actu et l'agrégateur 2424actu. Ceux-ci donnent accès à trois types de données :

– Les métadonnées
– Les documents
– Les requêtes

Ces données de départ sont de nature différente : texte court, texte long, texte transcrit, mesures, etc. Nous allons donc tenter de caractériser leurs spécifités qui vont par la suite se retrouver dans les corpus constitués.

5.1.1 Schéma du moteur et des données en présence

Le moteur de recherche de 2424actu est schématisé dans la figure 5.1. Le moteur fait intervenir les mêmes processus et données que dans le cas d'un modèle de RI classique (*cf.* figure 3.1). Les documents qui alimentent la base documentaire du moteur proviennent de sources variées, ils sont décrits en 5.1.3. Ils ont auparavant subi différents traitements automatisés (A) nécessaires aux regroupements de contenu pratiqués par l'agrégateur (clustering et thématisation décrites plus loin), ce qui permet de les enrichir avec un certain nombre de métadonnées (*cf.* 5.1.2). Cette base documentaire (B) est sauvegardée journalièrement.

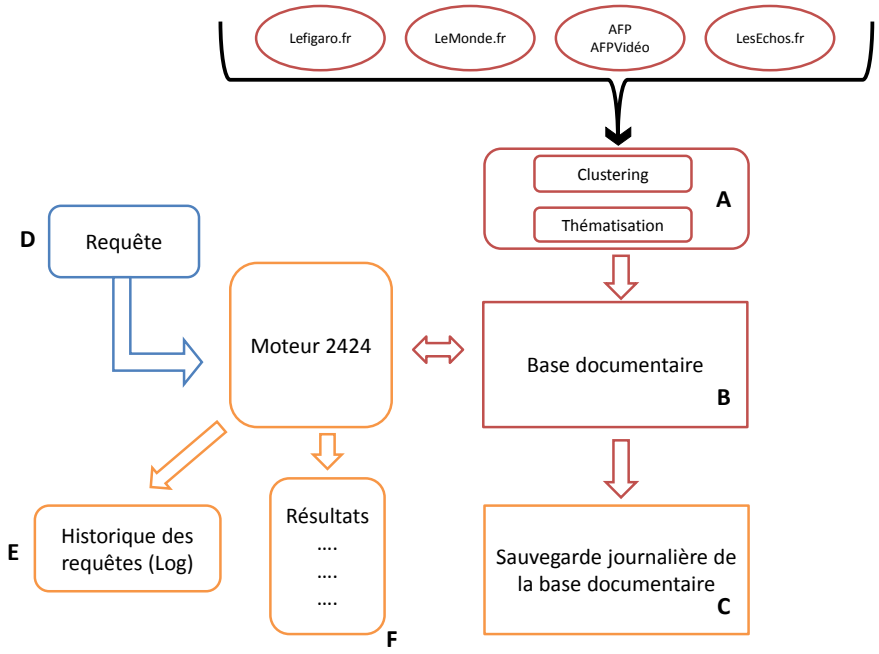


FIGURE 5.1 : Schéma du moteur 2424actu

D'autre part, les utilisateurs sont invités à soumettre leur requête au moteur 2424actu (D). Les requêtes soumises au moteur sont stockées dans un historique de recherche (E) aussi appelé *log* de requêtes que nous décrivons dans la sous-section 5.1.4. Les résultats (F) proposés en réponse à une requête sont ordonnés en fonction de leur similarité et de leur format distinguant les formats audio, vidéo et texte. Ces résultats ne sont pas accessibles pour des analyses ultérieures, contrairement à la base documentaire du moteur et aux historiques de recherche.

Deux modes de classification des documents

Les documents sont classifiés par deux techniques, ce qui permet deux niveaux de classement.

Le premier niveau de classification est la clusterisation des documents d'actualité provenant des fournisseurs (cf. 4.2.2). Elle permet de regrouper automatiquement les documents (audio, vidéo et texte) selon leur contenu. Les regroupements permettent de donner une image de l'actualité en temps réel. Ils sont par conséquent recalculés toutes les 20 minutes, afin d'intégrer les nouveaux flux. Ces regroupements sont opérés en s'appuyant sur les contenus provenant de l'AFP et en privilégiant les contenus de type vidéo. Les documents AFP servent en effet à ancrer la création d'un nouveau regroupement. Enfin, les contenus vidéos sont également privilégiés à cause du principe même de l'application 2424actu, qui propose un accès à l'actualité via le format vidéo (cf. 4.2.2).

Le deuxième niveau de classification est la thématisation des documents d'actualité. La thématisation repose sur l'étiquetage manuel *a priori* de certains documents fournis par les agences telles que l'AFP. Cet étiquetage est ensuite étendu automatiquement aux documents ne portant pas de thématique *a priori* grâce à la clusterisation réalisée auparavant : lorsqu'un document n'est pas étiqueté thématiquement dans un cluster, la catégorie thématique majoritaire lui est attribuée. Les documents vidéos ne sont pas étiquetés thématiquement. Les catégories sont au nombre de six : SOCIÉTÉ, INTERNATIONAL, POLITIQUE, ECONOMIE, SPORT, CULTURES. Ces catégories sont explicitées dans le tableau 5.2. Elles font partie des métadonnées associées à chaque document sauvegardé journalièrement. Ces métadonnées sont utilisées de manière privilégiée dans le chapitre 6.

Les catégories INTERNATIONAL, SOCIÉTÉ et POLITIQUE sont géolocalisées, c'est-à-dire qu'une restriction géographique s'applique à ces catégories. En effet, globalement ces catégories expriment un point de vue franco-centré.

Cependant, il peut arriver que les restrictions géographiques ne soient pas homogènes. En effet, la notion spatiale est envisagée depuis le point de vue du journaliste. Ainsi, si un quotidien francophone comme *la Nouvelle Afrique*, traite les changements politiques et sociétaux qui ont été opérés en Tunisie en 2011, il va les catégoriser en POLITIQUE ou même en SOCIÉTÉ.

Catégories thématiques	Définition des catégories
INTERNATIONAL	Actualités hors de France
SOCIÉTÉ	Actualités en France et faits de société
POLITIQUE	Informations concernant la politique française
ECONOMIE	Actualités économiques sans restriction géographique
SPORT	Actualités sur le sport et les manifestations sportives sans restriction géographique
CULTURES	Informations culturelles, vie des médias avec une forte composante presse dite « people »

TABLEAU 5.2 : Les catégories thématiques dans 2424actu

5.1.2 Les métadonnées

Les documents sauvegardés chaque jour ont connu des modifications suite à leur passage dans l'agrégateur. Des métadonnées sont associées à chaque document. Ces métadonnées cumulent en particulier : la présence d'une image, la nature du document, les résultats des traitements appliqués en amont du site 2424actu comme la classification des documents ou les thématiques AFP associées. Le tableau 5.3 rassemble les métadonnées disponibles pour un document donné. Il existe deux types de métadonnées :

– les métadonnées liées aux données (type A dans le tableau) : comme la « source du document » ou la « date de publication » qui sont des apports d'information livrés avec les données.
– les métadonnées générées à partir de traitement (type B) : comme les identifiants du cluster référent ou du document centroïde (au centre du cluster) proche du document considéré. Elles permettent de situer le document par rapport aux documents d'actualité à un instant t.

Une partie des métadonnées sont des métadonnées comme les identifiants du cluster référent ou du document centroïde (au centre du cluster) proche du document considéré. Elles permettent de situer le document par rapport aux documents d'actualité à un instant t. Les métadonnées du type « source

du document » ou « thématique » sont des apports d'information touchant au contenu en lui-même.

Type	Métadonnées	Exemple
A	Nom de la source du document	LESECHOS
	Type de document	text
	Date de publication	2011-08-11
	Identifiant image	1543.jpg
B	Identifiant du document	5703730
	Identifiant du cluster auquel le document appartient	3295
	Identifiant du document « centroïde » du cluster	5703616
	Thématique du cluster	int
	Thématique du document	int

TABLEAU 5.3 : Les principales métadonnées d'un document 2424actu

5.1.3 Le format des documents

L'ensemble des documents sont sous un format textuel, mais pour environ 20% d'entre eux, le format texte est issu d'une transcription. En effet, trois types de format coexistent : audio, vidéo et texte. Toutefois, tous ces formats ne sont conservés que sous une forme textuelle.

Presse écrite

Les documents écrits sont issus d'articles journalistiques comme dans l'exemple 5.4. Un document texte est constitué d'un titre et d'un texte court, souvent tronqué. On retrouve les métadonnées décrites en 5.3, le document proposé en exemple provient du journal Les Echos.fr est catégorisé en INTERNATIONAL.

Vidéo et audio

Les documents initialement aux formats *vidéo* et *audio* sont des documents issus d'une transcription automatique. Le contenu transcrit est beaucoup moins homogène qu'un document texte parce qu'il peut avoir subi des traitements (segmentation). Par ailleurs, comme on peut le voir dans l'exemple 5.5 qui provient d'une vidéo AFP catégorisée en SOCIÉTÉ, le document qui contient de l'oral transcrit comporte différentes erreurs : l'expression *les deux bouts* est transcrite à tort par le segment *les debout*. Le document vidéo peut être aussi

accompagné de descriptions concernant l'ambiance et le déroulement de la vidéo.

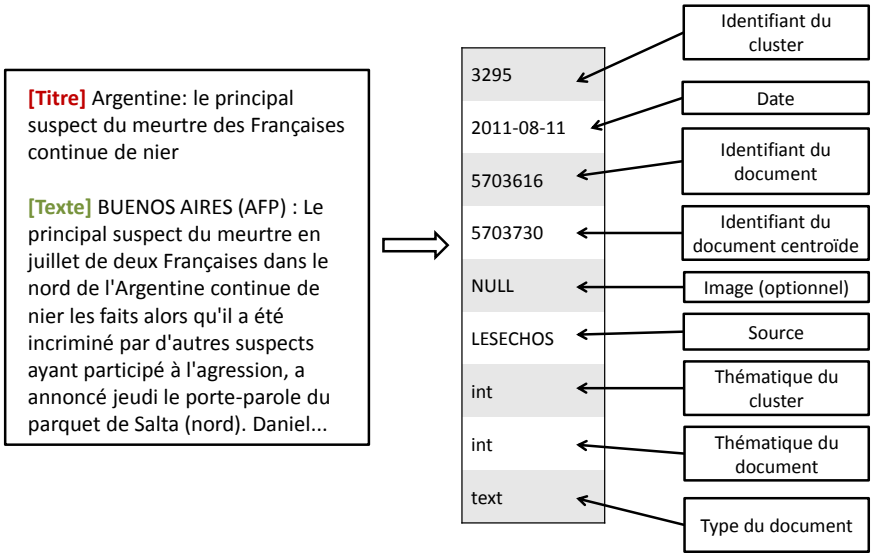


FIGURE 5.4 : Exemple d'un document sous forme de texte publié le 11 août 2011

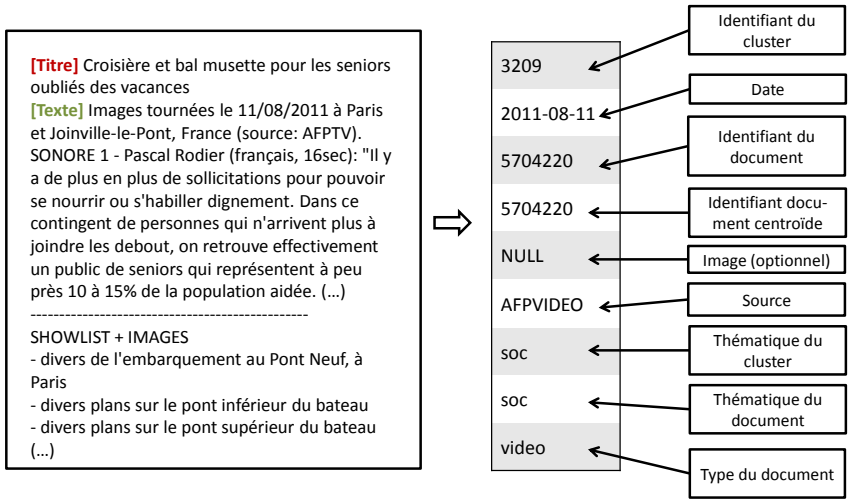


FIGURE 5.5 : Exemple d'un document vidéo publié le 11 août 2011

5.1.4 Le format des requêtes

Les requêtes des utilisateurs sont contenues dans l'historique du moteur de recherche sous la forme de *logs*. Nous avons vu dans le chapitre 3 que l'historique du moteur de recherche est utilisé pour accéder aux traces contextuelles disponibles. Un log est composé de plusieurs champs comme on peut le voir sur la figure 5.6. La date de la soumission de la requête compose le premier champ, la requête en elle-même est dans le deuxième champ. Du champ 3 à 5, est indiqué le nombre de documents retournés selon leur type : les documents audio (*fa*), les documents vidéos (*fv*) et les documents textuels (*ft*). Le champ 6 renseigne le type de support numérique (ordinateur, tablette, téléphone) qui a été utilisé pour envoyer la requête.

Date et heure de soumission	Requête	fa	fv	ft	support
[2011/08/21 09:15:56]	barack obama	22	103	692	pc
[2011/08/21 09:17:50]	baleine a marseillan	0	2	0	pc
[2011/08/21 09:27:56]	barack obama	22	103	692	pc
[2011/08/21 09:29:40]	libye	17	115	910	pc
[2011/08/21 09:30:54]	irlande france	28	6	226	pc
[2011/08/21 09:32:52]	deligones	0	0	0	pc
[2011/08/21 09:33:06]	deligones	0	0	0	pc

FIGURE 5.6 : Extrait du *log* de requêtes du moteur 2424actu le 8 août 2011

5.2 Le corpus constitué

A partir des données de départ fournies par le moteur et l'agrégateur 2424actu, nous avons constitué deux corpus de nature différente : un corpus de documents et un corpus de requêtes. Ces deux corpus sont alignés temporellement. Nous allons tout d'abord présenter la méthodologie qui a permis de constituer ces deux corpus, puis nous allons décrire ces corpus en fonction de leurs caractéristiques.

5.2.1 Les documents

Les documents ont été collectés de manière automatique toutes les nuits à heure fixe grâce à un programme informatique de sauvegarde [1]. La collecte a été soumise à certains aléas. Les principaux problèmes rencontrés ont été des

1. créé par Johannes Heinecke

interruptions imprévues du service 2424actu, et surtout l'arrêt de l'application qui a été amorcé dès le mois d'août 2011 pour devenir définitif au 1er octobre 2011. Il a fallu composer avec les arrivées de nouveaux fournisseurs ou les changements de format.

La constitution du corpus a été réalisée en deux étapes. Un premier travail a permis d'obtenir un corpus sous forme de base de données conservant les métadonnées, et permettant de réaliser les premiers filtres sur la langue des documents, et les doublons. La deuxième étape a consisté à transformer le corpus vers un format XML, compatible avec le moteur de recherche Terrier [2] (cf. 7.1.1.3). Cette étape nous a également permis de repérer par des phases d'aller-retour avec le moteur de recherche des bugs qui ont touché ponctuellement la génération automatique des identifiants des documents du site 2424actu.

Dans un premier temps, nous présentons la constitution du corpus et le travail de nettoyage qui a été réalisé. Dans un second temps, ce sont les caractéristiques quantitatives du corpus qui sont examinées.

5.2.1.1 Processus de nettoyage et de constitution du corpus

Un programme nous a permis de sauvegarder automatiquement chaque nuit les documents du jour. Les documents sont conservés pendant un mois. Cela permet d'agréger l'actualité de manière continue. Nous avons constitué au total trois corpus de documents, un premier corpus de pré-tests (*corpus 2424beta*), un deuxième corpus pour l'expérimentation (*corpus 2424)* et un troisième corpus est dédié à l'évaluation (*corpus 2424suite*).

Pour construire le corpus 2424beta, nous avons utilisé les sauvegardes du dernier jour de chaque mois. Cela nous a permis d'avoir les documents disponibles pour les utilisateurs de la période de Mai 2010 à Décembre 2010. Le travail de nettoyage était alors semi-automatique (passage de filtres pour éliminer les documents en anglais, les erreurs de formatage). Ce premier corpus a permis de faire les premières observations mais le nettoyage n'était pas satisfaisant, en particulier sur le filtrage des documents en langue anglaise. En effet, de nouveaux fournisseurs anglophones ont été intégrés pendant la collecte des données.

Nous avons par la suite développé une véritable chaîne de traitement exhaustive et modulable. En effet, il s'agissait également de s'adapter aux différents changements de format qui sont intervenus entre Mai 2010 et Août

2. http://terrier.org/

2011. Cette chaîne a permis de constituer deux corpus : le corpus 2424 et le corpus 2424suite.

La chaîne de traitement est composée de sous-programmes (développés en python) gérés par un programme bash. Pour traiter une nouvelle partition du corpus, il suffit d'indiquer le fichier à traiter lorsqu'on lance le programme bash et de préciser dans ce programme le nom du fichier de sortie. Le processus débute par la concaténation des fichiers de « backup » d'un mois en entier. Les documents présents plusieurs fois sont filtrés, tout comme les sources de médias d'actualité de langue anglaise. Nous avons décidé de ne garder qu'une seule thématique pour un document qui aurait changé de thématique plusieurs fois lors des calculs quotidiens des clusters [3]. Le choix s'est fait de manière arbitraire, la thématique retenue est celle qui est apparue en premier. Une conversion du corpus traité au format UTF-8 est également opérée. Les multiples métadonnées comme l'identifiant du cluster, de la news centrale du cluster ou les images sont filtrées pour ne retenir au final que les métadonnées suivantes : identifiant du document, date, source, thème, titre, texte.

Le corpus obtenu est au format XML, sous la forme d'un arbre XML pour chacun des documents, comme on peut le voir dans l'exemple 5.7 qui montre un extrait d'un document ainsi nettoyé. Cet exemple est un document édité par le journal *Tribune de Genève* relatif à des massacres perpétués en République Démocratique du Congo. Il est étiqueté en international (INT).

```
<DOC>
  <IDENT>1065787</IDENT>
  <DATE>2010-10-01</DATE>
  <SOURCE>TRIBUNEGENEVE</SOURCE>
  <THEME>int</THEME>
  <TITLE>ONU : le rapport sur les atrocités en RDC, un premier pas contre l'impunité</TITLE>
 <TEXT>La publication vendredi d'un rapport de l'ONU détaillant la mécanique de l'horreur en République démocratique du Congo (RDC) de 1993 à 2003 se veut, au-delà de la polémique sur d'éventuels crimes de génocide par l'armée rwandaise, un premier pas sans précédent pour venir à bout de l'impunité dans une région martyre. </TEXT>
</DOC>
```

FIGURE 5.7 : Le document au format XML

3. Les catégories thématiques dépendent du calcul des clusters (cf. 5.1.1). Si un document est réaffecté à un nouveau cluster (par exemple division d'un cluster devenu trop important en taille pour être pertinent), il peut avoir être re-thématisé. Or pour des raisons de simplicité, nous avons préféré ne garder qu'une thématique par document.

5.2.1.2 Description des corpus 2424

Les corpus sont décrits dans les tableaux 5.8, 5.9 et 5.10. Chaque sous-corpus correspond à un mois d'actualité.

Sous-corpus	Nombre de documents
Mai	23 521
Juin	26 782
Juillet	15 773
Aout	19 543
Septembre	17 634
Octobre	22 822
Novembre	16 015
Décembre	11 096
Total	153 190

TABLEAU 5.8 : Corpus 2424beta (2010)

Dans le corpus 2424 (tableau 5.9), les sous-corpus sont plutôt équilibrés en nombre de documents. Au total, le corpus constitué contient 152 772 documents, le nombre de mots est de 14 847 983. Le nombre de documents est inférieur au corpus 2424beta, le filtrage des documents d'actualité en anglais étant plus performant. Il est utilisé dans le chapitre 8 comme base documentaire du moteur sur lequel se sont déroulés les analyses de cooccurrences alors que le corpus 2424beta est utilisé par la première expérience sur la catégorisation (chapitre 6).

Sous-corpus	Nombre de documents	Nombre de mots
Mai	20 872	1 830 560
Juin	22 073	2 153 961
Juillet	13 412	1 125 317
Aout	24 031	2 084 088
Septembre	20 082	2 079 199
Octobre	20 391	2 036 111
Novembre	17 791	1 949 251
Décembre	14 120	1 589 496
Total	*152 772*	*14 847 983*

TABLEAU 5.9 : Statistiques du corpus 2424 (2010)

Le corpus 2424suite est constitué de deux sous-corpus : janvier et février 2011 (cf. tableau 5.10). Il comporte 44 140 documents au total. Il y a 4 732 242 mots dans le corpus d'évaluation, les deux sous-corpus présentant un nombre de mots équivalent. Ce corpus prolonge temporellement le corpus 2424. Il est utilisé dans le chapitre 7 comme base documentaire du moteur sur lequel se sont déroulés les tests utilisateurs.

Sous-corpus	Nombre de documents	Nombre de mots
Janvier	24 066	2 517 172
Février	20 074	2 215 070
Total	*44 140*	*4 732 242*

TABLEAU 5.10 : Statistiques du corpus 2424suite (2011)

5.2.2 Les requêtes

Le corpus de requêtes a été constitué à partir des historiques de recherche du moteur 2424 comme précisé en 5.1.1. Les *logs* de requêtes de 2424actu devaient être récupérés tous les deux mois avant qu'ils ne soient effacés. En effet, l'entreprise ne conserve pas les historiques des requêtes des utilisateurs au-delà de ces deux mois. Par conséquent, la récupération des logs de requêtes n'est due qu'à la bonne volonté de Bénédicte Cherbonnel, responsable du projet 2424actu. Nous l'avons sollicitée le plus régulièrement possible afin d'obtenir les requêtes des utilisateurs. Toutefois, tous les *logs* n'ont pu être récupérés. Par exemple, le mois de septembre 2010 est incomplet.

5.2.2.1 Processus de nettoyage et de constitution du corpus de requêtes

La transformation de l'historique du moteur de recherche en une liste de requêtes exploitable a été faite grâce à des outils de filtrage élémentaires (grep, sort). La figure 5.11 schématise cette opération. Le nettoyage consiste à garder les requêtes, lesquelles sont situées dans la deuxième colonne du *log* de requêtes, assorties d'une information temporelle (information contenue dans la première colonne). Les troisième, quatrième et cinquième colonnes correspondent au nombre de documents retournés par le moteur. Les documents sont distingués en fonction de leur type, respectivement audio, vidéo et texte. L'opération de nettoyage du corpus va consister à utiliser l'ensemble des logs de requêtes correspondant à une période donnée, à extraire les requêtes en elles-même puis à les ordonnancer en fonction de leur fréquence.

Il faut noter que les requêtes ont été désaccentuées lorsqu'elles ont été traitées par le moteur de recherche. Nous avons procédé manuellement à une réaccentuation des requêtes lors des expérimentations présentées dans les chapitres suivants.

[2010/10/22 09:13:53]	inondations	9	34	285	pc
[2010/10/22 09:14:48]	larry clark	11	5	93	pc
[2010/10/22 09:18:13]	greve	303	407	2440	pc
[2010/10/22 09:18:21]	greve	303	407	2440	pc
[2010/10/22 09:18:42]	greve	303	407	2440	pc
[2010/10/22 09:19:30]	colette renard	4	4	44	pc
[2010/10/22 09:20:48]	nice	52	34	351	pc

log

14358	greve
...	...
3801	larry clark
3775	petrole
3489	reforme retraites
3056	sport foot
2827	boue hongrie

Corpus

FIGURE 5.11 : Constitution du corpus de requêtes pour le mois d'octobre 2010

De la même manière que pour les corpus de documents, les corpus de requêtes sont composés d'un corpus dédié à l'expérimentation (*corpus de requêtes 2424*) et d'un corpus destiné à l'évaluation (*corpus de requêtes 2424suite*).

5.2.2.2 Description des corpus de requêtes

Du point du vue quantitatif, le corpus constitué pour les expérimentations contient 487 231 requêtes (*cf.* le tableau 5.12). Les requêtes sont organisées en 8 sous-corpus qui correspondent chacun à un mois donné. Cependant, il faut rappeler que le log de requêtes du mois de septembre n'est pas complet, nous disposons seulement des historiques du 1erseptembre et du 23 septembre au 30 septembre 2010, ceci explique que le sous-corpus de septembre 2010 ne contienne que 16 888 requêtes (*cf.* le tableau 5.12 page suivante). D'autre part, on observe que les sous-corpus de juillet et août 2010 contiennent environ

deux fois moins de requêtes que les autres sous-corpus. Ceci est certainement dû à la baisse de fréquentation du site d'actualité en période de vacances.

Sous-corpus	Nombre de requêtes	Nombre de requêtes uniques
Mai	90 041	3 935
Juin	72 596	4 875
Juillet	43 064	4 219
Aout	43 547	4 955
Septembre	16 888	1 984
Octobre	89 066	7 140
Novembre	51 906	5 269
Décembre	80 123	6 489
Total	487 231	38 866

TABLEAU 5.12 : Statistiques du corpus de requêtes 2424 (2010)

Le corpus de requêtes 2424suite est composé seulement de deux sous-corpus : janvier et février 2011 (tableau 5.13). Il prolonge temporellement les sous-corpus du corpus de requêtes 2424 (tableau 5.12). Quantitativement, ce corpus d'évaluation contient 133 207 requêtes pour 14 582 requêtes uniques. De manière similaire au corpus 2424suite (voir tableau 5.10), ce corpus est utilisé dans le chapitre 7.

Sous-corpus	Nombre de requêtes	Nombre de requêtes uniques
Janvier	62 262	6 910
Fevrier	70 945	7 672
Total	133 207	14 582

TABLEAU 5.13 : Statistiques du corpus de requêtes 2424suite (2011)

5.3 Les contraintes applicatives et les avantages des données réelles

Le service a été conçu de manière à donner accès à l'actualité en temps réel avec une fenêtre temporelle réduite. Il ne donne pas la possibilité à l'utilisateur de rechercher des informations datées de plusieurs mois. Pour l'analyse des données, aucun dispositif n'est prévu pour collecter les contenus « périmés ». D'autre part, l'accès aux documents et aux requêtes des utilisateurs s'opère par des voies différentes (*cf.* 5.1). Nous avons donc mis en place un double processus de récupération des données, qui privilégie les données « complètes »,

c'est-à-dire les documents et les requêtes issus d'une même zone temporelle, permettant de suivre l'actualité et le comportement des utilisateurs sur plusieurs mois.

Le service une fois déployé n'a pas été conçu pour être un lieu d'expérimentations, le moteur mis à disposition des usagers n'était donc pas accessible. En effet, l'application 2424actu est conçue de telle manière que tous les composants techniques ne sont pas localisés au même endroit, multipliant le nombre d'interlocuteurs. Cela a été une contrainte forte sur nos travaux. Les informations contextuelles telles que les documents retournés à l'utilisateur ou l'identifiant unique de chaque utilisateur ne sont également pas disponibles. La récupération de telles données n'était pas prévu initialement par le projet et aucune modification n'était possible. Une grande partie de notre travail a consisté à retrouver des conditions d'expérimentation satisfaisantes en particulier en utilisant et paramétrant un moteur de recherche libre d'accès, Terrier (*cf.* 7.1.1.3).

Les données issues de 2424actu sont atypiques par leur richesse et par leur empan temporel. En effet, les données recueillies présentent un volume intéressant et une grande diversité, à la fois en terme de format (texte et oral transcrit) et de diversité journalistique (plus de 50 fournisseurs de contenus d'actualité différents). C'est un type de données peu étudié dans les travaux de l'état de l'art. En effet, les corpus de requêtes étudiés sont la plupart du temps issus de campagne d'évaluation comme TREC [4] (base documentaire constituée de journaux) et moins couramment d'un contexte commercial en accès limité.

La temporalité des données est également un élément clé. Au total, 11 mois de requêtes et de documents ont été récupérés et alignés. Ce suivi longitudinal permet de reconstituer le contexte de l'application mais aussi le contexte des utilisateurs de cette application. La temporalité permet ainsi de prendre en compte la variable temporelle portée par l'actualité en elle-même et la dimension contextuelle du processus de recherche.

Enfin, nos données sont entièrement anonymisées. Ce fait indépendant de notre volonté empêche toute intrusion dans la vie privée des utilisateurs de l'application 2424actu, ce qui est un élément positif. Toutefois, l'anonymisation des données élimine toute trace de l'utilisation du contexte du processus de recherche, rendant par exemple impossible la délimitation des sessions de recherche de chaque utilisateur.

4. Voir le chapitre 2, section 2.2.2.2

5.4 Pratiques de recherche dans l'actualité

Nous amorçons ici une première caractérisation des requêtes présentes dans le corpus 2424actu selon différents aspects. Cette caractérisation permet de percevoir les spécificités des requêtes formulées par les utilisateurs. Ces requêtes peuvent en effet être considérées comme la marque, la trace des pratiques de recherche des utilisateurs. L'analyse de ces traces apporte un certain nombre d'informations précieuses sur le contexte du processus de recherche et sur le contexte utilisateur en lui-même. Enfin, comme nous l'avons vu, l'information temporelle est un élément du contexte de l'application accessible dans le corpus de requêtes 2424 (cf. section 4.3.2). Nous avons utilisé cette information pour étudier les requêtes utilisateurs sous un angle particulier, celui du profil temporel. Enfin, nous avons caractérisé les requêtes du point de vue linguistique, en mettant en évidence l'importance des entités nommées dans le contexte de recherche sur l'actualité. Nous retraçons ces différentes étapes dans ce qui suit.

5.4.1 Première caractérisation des requêtes

Une manière de caractériser les pratiques de recherche des utilisateurs est de typer les requêtes du point de vue de leur fonction. Rappelons [5] que la classification de Broder (2002) distingue trois catégories de requêtes : navigationnelle, transactionnelle et informationnelle.

Une requête navigationnelle traduit l'intention immédiate d'un utilisateur à se rendre sur un site qu'il spécifie. Elle se traduit par des requêtes relativement courtes, de trois termes en moyenne (Jansen *et al.*, 2007). Ce type de requête présente des caractéristiques particulières telles que la présence de :
– noms de compagnie, d'organisation ;
– noms de domaines.
Les requêtes dites « transactionnelles » fonctionnent comme des intermédiaires entre le moteur de recherche et une activité commerciale. Selon Jansen *et al.* (2007), le contenu sémantique de la requête ou la présence de certains termes comme *obtenir*, *télécharger, audio, images, chansons* permet de les caractériser. Les requêtes navigationnelles sont les plus simples à identifier car elles comportent la plupart du temps des adresses internet complètes ou semi-complètes (Jansen *et al.*, 2007).

Une requête informationnelle montre que l'utilisateur souhaite trouver et acquérir une information (Broder, 2002). La caractérisation de ce type de requête

5. Voir la section 3.2.2.2.

est difficile : Jansen *et al.* (2007) indiquent que les requêtes informationnelles peuvent contenir des pronoms interrogatifs (*qui*, *quoi*) ou des adverbes (*combien*, *comment*). Ce type de requête est donc défini de manière négative par rapport aux deux autres types de requêtes. Le type informationnel se caractérise par l'absence des traits indiquant une requête navigationnelle ou transactionnelle.

Afin de faire émerger les caractéristiques spécifiques aux requêtes de 2424actu, il est intéressant de les confrontrer à des requêtes d'un moteur généraliste (cf. 4.1), le moteur du Portail Orange [6]. Dans ce but, nous proposons de comparer les requêtes les plus fréquentes du mois d'octobre 2010 soumises au moteur 2424actu et les requêtes les plus fréquentes du mois de janvier 2010 provenant du moteur du Portail Orange. La table 5.15 illustre la comparaison menée entre les deux corpus de requêtes.

Rang	Requêtes 2424 (10/2010)	Rang	Requêtes Orange (01/2010)
1	*greve* (14 358)	1	*google* (9 258 260) N
2	*greve rer* (12 179)	2	*_google_monitor_query_ or googletestad* (3 814 367)
3	*mineurs chili fr* (5 549)	3	*facebook* (2 549 477)
4	*afghanistan* (3 978)	4	*le bon coin* (1 032 810)
5	*larry clark* (3 801)	5	*youtube* (735 954)
6	*petrole* (3 775)	6	*yahoo* (579 452)
7	*reforme retraites* (3 489)	7	*leboncoin* (574 268)
8	*sport foot* (3 056)	8	*ebay* (454 646)
9	*boue hongrie* (2 827)	9	*pages jaunes* (422 502)
10	*thailande* (1 913)	10	*you tube* (387 745)

TABLEAU 5.15 : Comparaison des requêtes les plus fréquentes du moteur 2424 actu et celui de Portail Orange - Fréquence des requêtes.

On note tout d'abord que les fréquences des requêtes des deux corpus ne sont pas comparables, le moteur généraliste mobilisant un traffic bien supérieur à celui du site d'actualité. Par ailleurs, on constate un décalage entre la requête la plus fréquente et les suivantes : *greve* sur 2424 est 7 fois plus fréquente que la requête *thailande* (rang 10). L'écart est beaucoup plus prononcé pour la requête *google* qui est 23 fois plus fréquente que la requête *you tube* au rang 10. On constate un effet particulier sur le moteur du Portail Orange, il est fortement

6. http://www.orange.fr/portail

sollicité par les utilisateurs pour rejoindre un site, en l'occurrence le moteur de recherche Google.

On observe dans la table 5.15 de très claires différences entre les requêtes des deux moteurs. Si les utilisateurs de 2424actu produisent majoritairement des requêtes informationnelles, les utilisateurs de Portail Orange formulent des requêtes navigationnelles qui cherchent à atteindre un autre site : *le bon coin, leboncoin, ebay*. On note la présence au rang 2 des requêtes les plus fréquentes du moteur Portail Orange d'une requête provenant d'un robot. À noter que les deux premières requêtes sont des variantes d'une même requête. Le moteur généraliste est donc largement utilisé comme un moyen d'accéder à des sites fréquemment consultés à la manière d'une liste des sites « favoris » de l'utilisateur.

A l'inverse, les requêtes 2424 sont clairement informationnelles. On remarque deux requêtes composées de noms de pays : *afghanistan* et *thailande*. Quatre requêtes sont formées par juxtaposition de mots : *reforme retraites*, *sport foot*, *mineurs chili fr* et *boue hongrie*. Ces requêtes visent des évènements particuliers, elles sont donc informationnelles. Par exemple, la requête *sport foot* est intéressante parce qu'elle intègre une spécification (*sport*). Celle-ci est recommandée lorsqu'on cherche des informations sur le foot dans un moteur généraliste et en anglais où il y a une concurrence entre la mesure et le sport. Nous remarquons également la requête *petrole* dont l'intention informationnelle est certainement à rapprocher des deux requêtes les plus populaires au mois d'octobre 2010 : *grève* et *grève rer*. En effet, ces requêtes informationnelles et typique du domaine de l'actualité nous informent sur la présence du mouvement social contre la réforme des retraites à cette période donnée. Les raffineries ont été bloquées pendant de nombreux jours ce qui a fait craindre une pénurie de pétrole. Nous notons également que *greve rer* est une extension de la requête *greve*.

Enfin, plusieurs de ces requêtes font appel à la notion d'évènement, comme la requête *larry clark,* qui ne désigne non pas le nom du photographe et réalisateur américain, mais une retrospective de ses œuvres a eu lieu en octobre 2010 à Paris. C'est ce que l'on retrouve dans la requête *boue hongrie* qui réfère à une catastrophe climatique ayant touchée la Hongrie en octobre 2010. Nous avons vu dans le chapitre 1 que les noms propres peuvent prendre des valeurs événementielles (Lecolle, 2004; Gary-Prieur, 2001). Dans le cadre particulier de la presse, on les appelle aussi des « noms propres d'évènements » (Krieg-Planque, 2009). Ces formes linguistiques sont très utilisées par les journalistes pour promouvoir une occurrence désignant un évènement perçu comme visible et symptomatique de ce qui se passe dans l'espace public (Krieg-Planque,

2009). Ce sont des occurrences qui s'interprètent seulement en contexte, dans un cadre particulier.

Les deux moteurs présentent donc des usages différents et les utilisateurs les sollicitent pour des raisons différentes. Le moteur du Portail Orange est utilisé majoritairement pour atteindre un site recherché à l'extérieur du portail. L'utilisateur de 2424actu recherche des informations propres à l'actualité, et, à la différence de l'utilisateur du moteur généraliste, il ne cherche pas à sortir du site 2424actu mais à obtenir du contenu proposé par l'agrégateur d'actualité.

5.4.2 Taille des requêtes

La taille moyenne d'une requête est un premier critère porteur de renseignement sur le degré de spécificité de l'information exprimée. L'étude de Spink *et al.* (2002b) a ainsi montré que la longueur des requêtes en anglais (provenant du moteur Excite Web) était de 2,6 mots en moyenne et qu'elle variait peu dans le temps. Dans le corpus 2424actu, la longueur moyenne d'une requête est encore inférieure, puisqu'elle est de 1,73 mots. On constate par ailleurs dans la figure 5.16 que la longueur des requêtes tend à diminuer dans notre corpus au fil des mois. Cette diminution est certainement à attribuer au déploiement de l'application mobile de 2424actu à partir de septembre 2010, mais n'ayant pas de chiffres précis à ce propos, il n'est pas possible de l'affirmer.

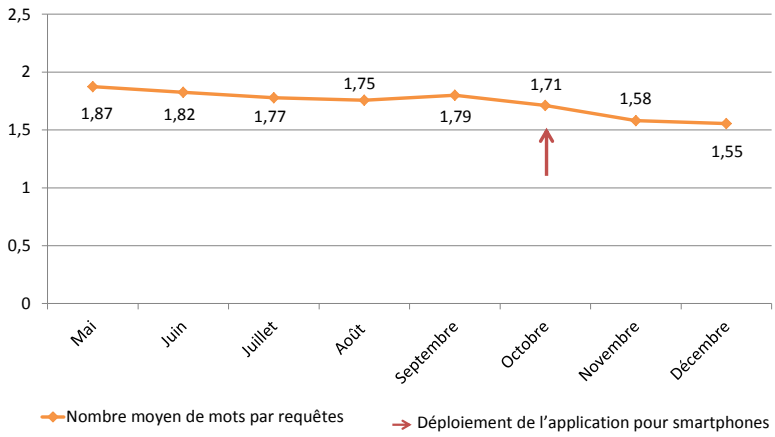


FIGURE 5.16 : Nombre moyen de mots par requête dans le corpus 2424actu

Dans notre corpus de requêtes, la longueur moyenne d'une requête est donc inférieure à 2 mots. Lorsqu'on a affaire à des requêtes multi-mots, ce sont des

termes nominaux complexes ou des noms propres composés comme *assemblée nationale* ou *côte d'ivoire*. Les autres termes nominaux complexes sont formés par l'effacement d'un joncteur grammatical comme pour les requêtes *réforme retraites* ou *grève RER*. Parmi ces requêtes composées de termes complexes, on distingue des requêtes associant un terme à une spécification d'ordre temporel ou spatial (*boue hongrie*, *marée noire etats-unis*).

L'étude de Barr *et al.* (2008) montre que les requêtes Yahoo ! en anglais sont composées en grande partie de noms propres (40%) et de noms communs (30%). Pour cela, ils ont procédé à une annotation manuelle des requêtes. Mais ce genre d'étude est tout de même difficile à mener. En effet, les mots présents dans les requêtes sont souvent ambigus du point de vue catégoriel, comme par exemple *chatting* qui peut être considéré comme un verbe ou un gérondif, ou *download* qui peut être un verbe ou nom. Ce type d'ambiguïté souffre de l'absence de contexte et rend difficile une tâche d'étiquetage automatique, et même manuel.

Le corpus des requêtes les plus fréquentes de 2424actu Afin de réaliser des observations plus poussées du corpus des requêtes de 2424actu, nous avons constitué un corpus échantillonnant le corpus général de 2424actu (cf. tableau 5.12). Dans ce but, nous avons choisi de garder les 49 requêtes les plus fréquentes de chaque sous-corpus (chaque mois) du corpus 2424actu. Ce sont donc des requêtes populaires tel que défini en 3.2.3.1. Ce corpus contient donc 391 requêtes[7]. Désormais, ce corpus sera le *corpus 2424reqFréquentes*. Le corpus est visible en Annexe A.2.

5.4.3 La place des entités nommées

La notion d'entité nommée (EN) a été créée par la communauté du TAL. Les EN sont majoritairement des noms propres (personnes, lieux, organisations) comme *François Hollande,* ou des descriptions définies (*le premier ministre*), mais aussi des expressions numériques et temporelles comme *11 septembre* ou *20h20* (Ehrmann, 2008). Les entités servent d'« ancres référentielles dans le flux informationnel » (Nazarenko, 2005). Cette fonction les rend intéressantes à repérer dans un document. Les EN permettent donc de repérer les référents d'un discours. Par exemple, dans un titre d'article d'actualité comme « ONU : le rapport sur les atrocités en RDC, un premier pas contre l'impunité », on

7. Suite à la première expérience (chapitre 6), le sous-corpus du mois de décembre a été amputé d'une requête.

repère deux entités *ONU* et *RDC* ; on peut donc dire que l'article parle de la République Démocratique du Congo et que l'ONU est impliqué.

Cependant, comme le dit Nazarenko (2005), si les EN sont facilement repérables en contexte, identifier le référent d'une entité nommée n'est pas toujours immédiat.

Nous avons étudié la présence et la répartition des EN dans les requêtes 2424-actu. Le repérage des EN s'appuie avant tout sur des ressources et sur le contexte qui les entoure. En l'absence de contexte et sur un corpus réduit, nous avons préféré procéder à une annotation manuelle des EN dans les requêtes. Nous avons effectué cette annotation sur le corpus 2424reqFréquentes.

Concernant les types d'EN, les EN annotées regroupent des noms propres de personne, des noms de lieux et des noms d'organisation. Les noms de personne peuvent être composés seulement d'un nom de famille comme *sarkozy* ou bien comportés un nom et un prénom comme *larry clark.* Les noms de lieux sont principalement des noms de pays (*afghanistan*, *thailande*) et des noms de ville (*lyon, toulouse*). Les noms d'organisation désignent des noms d'entreprise comme *airbus* ou *facebook*. Les expressions numériques ou temporelles ne sont pas présentes dans les requêtes fréquentes de 2424actu.

Concernant leur fréquence, il s'avère que la présence des entités nommées est importante parmi les requêtes fréquentes de chaque sous-corpus. En effet, 69% de requêtes de ce corpus contiennent ou sont une EN. Il y a également en moyenne 33 requêtes composées d'EN par sous-corpus. Cette présence se renforce considérablement dans la seconde moitié des sous-corpus allant de septembre à décembre. Cette augmentation des EN dans les requêtes est certainement à rapprocher de la baisse de la longueur moyenne des requêtes (cf. figure 5.16). En effet, les EN sont le plus souvent utilisées seules dans une requête et elles comportent un ou deux mots en général.

5.4.4 Profils temporels des requêtes

L'organisation temporelle du corpus de requêtes permet de faire émerger une caractéristique contextuelle importante d'un système de RI, l'évolution temporelle des requêtes des utilisateurs, comme on l'a vu précédemment (section 3.2.3.1). De manière générale, les informations temporelles ont principalement été utilisées pour la détection de sujets et le suivi de ces sujets (Swan et Jensen, 2000). Ces méthodes ont inspiré la classification temporelle des requêtes, effectuée à partir de la distribution des requêtes dans les documents. Par exemple, Diaz et Jones (2004) ont évalué le niveau de complexité d'une requête en intégrant des profils temporels. Le profil temporel d'une requête reconstitue la

distribution de celle-ci dans les documents d'une collection. Il est calculé à partir de l'apparition de la requête considérée dans une collection de documents organisée temporellement, en l'occurrence des articles de journaux. Ce type de profil permet par exemple d'observer qu'une requête comme *earthquake in armenia* (tremblement de terre en arménie) apparaît à une période particulière du corpus étudié, sa fréquence d'apparition formant un pic à ce moment là. Néanmoins, les auteurs ont analysé des requêtes provenant des campagnes d'évaluation TREC qui ne proviennent pas d'un historique de recherche. Ils ont analysé la manière dont se distribue une requête dans la collection de documents organisée chronologiquement et non quand une requête est formulée. Nous proposons de nous intéresser seulement à la réalisation des requêtes du point de vue temporel afin de rendre compte du comportement utilisateur et de la fluctuation de l'actualité.

Pour étudier les profils temporels des requêtes 2424actu, nous avons travaillé à partir du corpus 2424reqFréquentes décrit en 5.4.2. L'idée est de déterminer la durée de vie d'une requête dans ce corpus, soit entre mai 2010 et décembre 2010. La durée de vie correspond alors à l'empan temporel pendant lequel la requête est formulée au moins x fois. Ainsi, on peut savoir combien de fois une requête est apparue parmi les requêtes les plus fréquentes. Le profil temporel est donc basé sur deux informations : la date de production de la requête et la fréquence de celle-ci.

Pour des besoins de visualisation, les fréquences des requêtes sont ramenées à des fréquences relatives dans les graphiques suivants. Cela permet de diminuer l'effet provoqué par un sous-corpus incomplet. Celles-ci sont obtenues par la formule suivante qui permet d'obtenir la fréquence relative d'une requête, avec $f_{requête}$ qui correspond à la fréquence de la requête dans le corpus considéré et f_{totale} qui correspond au nombre total de requêtes du corpus considéré.

$$fréquence\,relative = \left[\frac{f_{requête}}{f_{totale}} \times 100\right] \tag{5.1}$$

Le calcul de la durée de vie des requêtes du corpus nous permet de distinguer deux profils de requêtes : les requêtes ponctuelles et les requêtes durables. Les requêtes ponctuelles sont soumises au moteur sur une durée limitée, que nous avons fixée de 1 à 4 mois pas nécessairement consécutifs. Les requêtes durables sont des requêtes qui sont fréquemment et régulièrement soumises par les utilisateurs au moteur de 2424actu. Ces deux types de profils sont illustrés dans la figure 5.17. On observe que les requêtes *greve* et *sarkozy* sont ponctuelles en comparaison des requêtes *afghanistan* et *haiti* qui ont été produites

continûment par les utilisateurs. Nous présentons un focus sur les deux types de profils de requêtes dans ce qui suit.

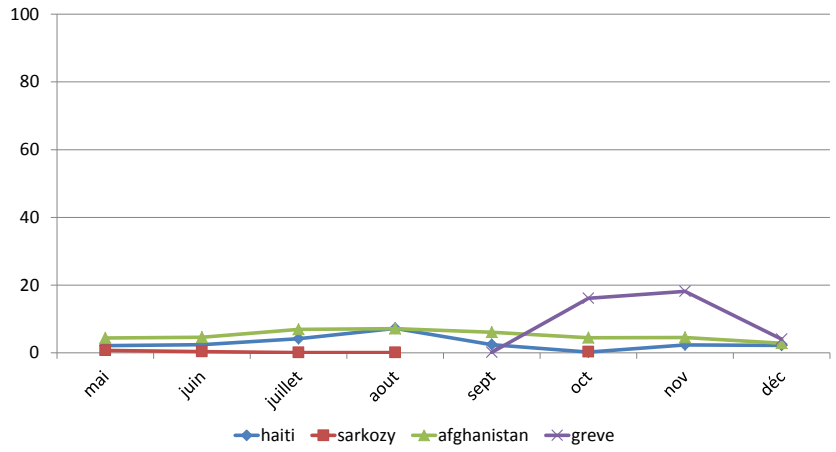


FIGURE 5.17 : Requêtes durables versus requêtes ponctuelles (fréquence relative)

Focus sur les requêtes ponctuelles Les requêtes ponctuelles sont les plus répandues parmi les requêtes du corpus *2424reqFréquentes*. Sont considérées comme « ponctuelles » les requêtes ayant une durée de vie inférieure ou égale à 4 mois. 91% des requêtes du corpus étudié répondent à ce critère. On peut observer quelques exemples de ces requêtes dans la figure 5.18. Les requêtes désignent des personnes (Jean-Luc Delarue, Laurent Fignon, etc.) mais également des évènements liés à leur personne (Larry Clark). Ainsi, la présence des requêtes comme *laurent fignon* est due au décès de Laurent Fignon (le 31/10/2010). La requête *larry clark*, qui désigne le photographe américain est liée à l'exposition qui a eu lieu au musée d'art moderne de Paris : cette exposition a été interdite aux moins de 18 ans ce qui a provoqué une polémique. La requête *delarue*, désigne également une personne et un évènement, en l'occurrence un fait-divers. On peut faire l'hypothèse, qui se vérifie sur ces quelques exemples, que les requêtes ayant une durée de vie ponctuelle ciblent des évènements marquants et délimités dans le temps. Parmi ces requêtes, on trouve *maree noire etats unis*.

Parmi ces requêtes ponctuelles, 51% des requêtes totales ont une durée de vie égale à un mois, c'est-à-dire qu'elles sont entrées une seule fois dans les 49 requêtes les plus fréquentes entre mai 2010 et décembre 2010. Ce pourcen-

tage est le signe d'un fort renouvellement des requêtes soumises au moteur 2424actu.

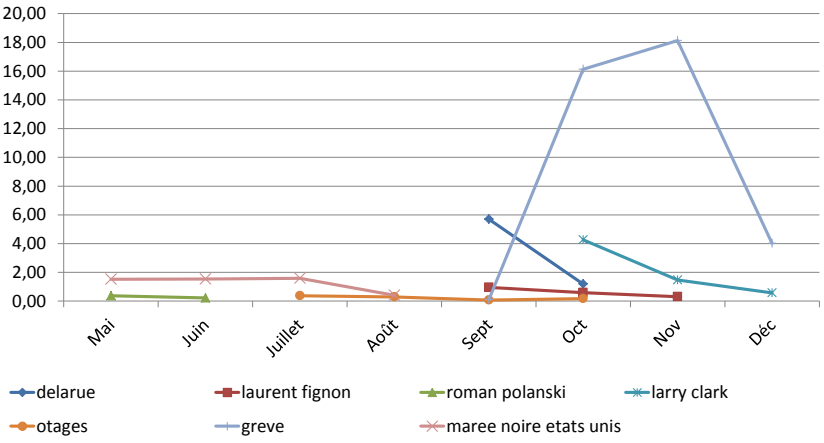


FIGURE 5.18 : Exemples de profils de requêtes « ponctuelles » (2424actu, fréquence relative)

Focus sur les requêtes durables On considère que les requêtes durables ont une durée de vie supérieure ou égale à 5 mois (le maximum étant 8 mois). Seulement 9 % des requêtes du corpus répondent à ce critère. Les requêtes durables sont essentiellement des noms de pays comme *afghanistan* (8 mois), *haiti* (8 mois), *israel* (8 mois), *pakistan* (8 mois) ou *thailande* (7 mois). Les profils temporels sont comparables pour les requêtes figurées en 5.19 mais pas les fréquences de ces requêtes. En effet, la requête *afghanistan* est plus fréquente que les autres requêtes.

La capacité de ces requêtes à être durable dans le temps interroge sur l'usage que les utilisateurs en font. Cette capacité semble être le signe d'une malléabilité potentielle de ces mots-clés. En effet, ces noms de pays ont pu être utilisés par les usagers du moteur pour accéder à des informations différentes au cours de cette période. Nous supposons en fait qu'un même mot-clé ne peut être durablement associé à une même information dans un contexte aussi volatile que l'actualité.

Les durées de vie non continues Jusqu'à présent, nous avons considéré les requêtes seulement sous l'angle de leur durée de vie en nombre de mois de la requête. Globalement, dans notre corpus, seulement 12 % des requêtes sont non continues. Des exemples de requêtes non continues sont visibles dans la

figure 5.20. Si certaines requêtes peuvent référer à des évènements ponctuels comme *woerth bettencourt*, d'autres réfèrent à des évènements obéissant à un calendrier particulier comme *sport tennis* ou *sport rugby*. On note également la présence de requêtes renvoyant à des personnes (*carla bruni*, *obama* ou *sarkozy*) qui ont une actualité régulière. Cependant, ces requêtes peuvent être touchées par des effets de seuils en étant par exemple moins fréquentes pendant une période particulière.

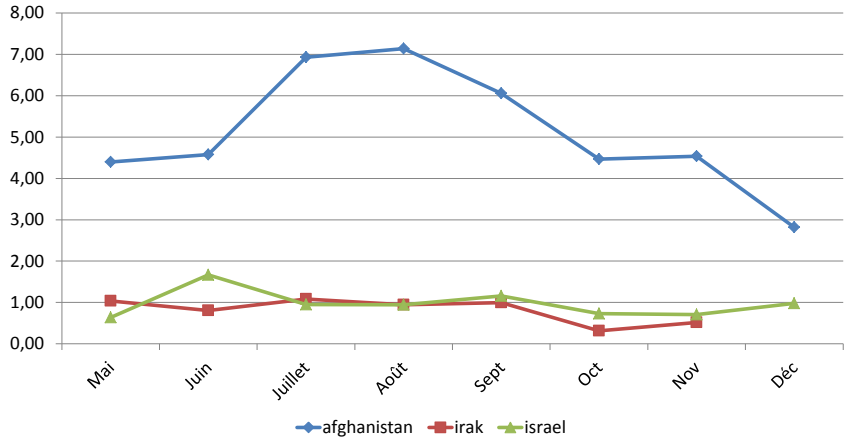


FIGURE 5.19 : Exemples de profils de requêtes « durables » (2424actu, fréquence relative)

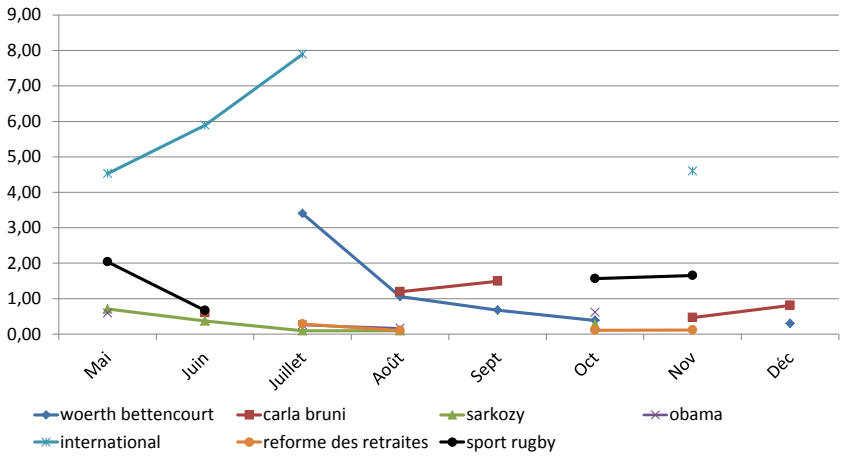


FIGURE 5.20 : Exemples de profils de requêtes avec une durée non continue (fréquence relative)

Conclusion La durée de vie des requêtes très fréquentes dans notre coprus laisse apparaître des comportements intéressants. En effet, l'étude temporelle nous a permis d'observer que le taux de renouvellement des requêtes de notre corpus est élevé (50% environ). C'est un résultat qui semble cohérent avec le domaine de l'actualité qui connaît un fort renouvellement. En ce sens, ce sont les requêtes durables qui génèrent des questionnements. Ce ne sont essentiellement que des noms de pays et on peut supposer que ce sont des requêtes malléables et potentiellement polyvalentes.

5.5 Conclusion

Le corpus constitué témoigne des avantages et des inconvénients d'un travail de recherche effectué dans un contexte industriel. Ce corpus volumineux permet de prendre en compte la diversité de l'actualité en ligne. Il offre également la possibilité d'observer la multitude de points de vue exprimés dans la presse en ligne à un instant donné. Le corpus de requêtes présente une grande variété dans les types d'informations recherchées. Les requêtes sont de temporalité variable. Toutefois, il a fallu pallier un certain nombre de problèmes comme la fragmentation des sources des données et l'inaccessibilité du moteur de recherche à proprement parler.

La caractérisation des requêtes issues du corpus 2424actu montre leur différence vis-à-vis des requêtes issues d'autres moteurs, en particulier de moteurs généralistes. Nous retenons cette prépondérance de requêtes informationnelles, ce type de requêtes est à prendre en compte lorsqu'on met en place des dispositifs d'interprétation de requêtes.

L'étude des profils temporels de requêtes montre la diversité des profils, avec deux types de distinction. Les requêtes durables méritent une attention particulière. En effet, il est difficile de supposer que dans un contexte où l'actualité est si volatile, un même mot-clé soit durablement associé à une information particulière. La durabilité d'une requête pourrait alors signaler une certaine capacité à la malléabilité et à l'adaptation au contexte.

La prépondérance des entités nommées et le fort taux de renouvellement des requêtes fréquentes posent la question du repérage et du traitement des requêtes, celles-ci apparaissent au gré de l'actualité et ne sont pas recensées par avance. Nous constatons également que les caractéristiques linguistiques des requêtes semblent être liées aux variations propres à l'application comme un déploiement au niveau mobile, montrant que les utilisateurs s'adaptent aux variations du contexte.

Enfin, la constitution et la caractérisation de nos données nous a permis de dégager un certain nombre d'éléments contextuels à notre disposition. Ce sont les catégories thématiques (métadonnées), les documents de la base documentaire et le log de requêtes. Ce sont sur ces éléments exploitables que s'appuie notre démarche d'expérimentation. Ces premiers éléments contextuels sont exploités dans les chapitres 6 et 7. Nous avons également identifié comme éléments contextuels, le contexte temporel porté par les requêtes, le contexte informationnel contenu dans les documents comme les cooccurrences et la capacité d'une requête à être spécifiée ou étendue. Ces éléments seront développés dans le chapitre 8.

Troisième partie

Émergence de l'ambiguïté des requêtes grâce à des indices contextuels

Chapitre 6

Un indice contextuel : la catégorisation thématique

Nous cherchons à identifier la présence de l'ambiguïté et la forme qu'elle peut prendre dans les requêtes que nous avons recueillies. Notre hypothèse est que la dispersion des résultats d'une requête peut être le signe d'une ambiguïté potentielle de la requête. Pour cela, nous souhaitons rendre compte d'une dispersion dans les résultats ramenés par une requête. Cette dispersion peut être matérialisée par deux types de méthodes : le clustering et la catégorisation. Nous avons choisi la catégorisation pour une raison principale. Les systèmes à partir de catégories sont plus faciles à utiliser pour les utilisateurs (Hearst, 2009). Les utilisateurs préfèrent en effet les hiérarchies dont on peut comprendre le sens et dont la granularité est uniforme (Hearst, 2011). En cela, la catégorisation présente un avantage indéniable puisque c'est un dispositif qui produit une hiérarchie réduite et fixe, contrairement au clustering.

Dans ce but, nous allons nous appuyer sur un indice contextuel à notre disposition : la catégorisation thématique. Il présente un intérêt certain pour nos recherches car c'est un mode de classement utilisé par les experts du domaine mais aussi un dispositif familier des utilisateurs, popularisé par les journaux et les sites d'actualité en ligne.

Ce chapitre se structure en deux temps. Dans un premier temps, nous mettons en place une expérience de catégorisation thématique des requêtes, et ce, en essayant de contrôler tous les biais possibles de cette méthode. Puis dans un second temps, nous testons la capacité du processus de catégorisation à révéler une diversité d'interprétations possibles pour une requête donnée. Ce test consiste à confrontrer la catégorisation thématique à une autre méthode de classification, Wikipédia. Ces deux méthodes donnent des résultats très

différents ce qui nous permet de comprendre l'apport spécifique de la catégorisation.

6.1 La catégorisation thématique des requêtes

La mise en place de la catégorisation thématique va se dérouler en plusieurs temps. Nous débutons par une présentation des hypothèses, puis nous exposons la méthode choisie pour procéder à la catégorisation et les résultats obtenus. Dans un second temps, nous examinons deux facteurs qui pourraient influencer la catégorisation des requêtes : la fréquence de la requête dans les documents et la distribution des catégories thématiques utilisées.

6.1.1 Hypothèses

Cette étude utilise les moyens disponibles dans le cadre de cette application, une double catégorisation des documents cibles des requêtes. Pour cela, nous nous servons de documents qui ont la particularité d'être classifiés à deux niveaux : chaque document (ou news) appartient à un regroupement de type sémantique (cluster) et à une catégorie thématique (cf. 5.1.1). Notre but est de parvenir à transposer une catégorisation externe (des documents) à une catégorisation interne (des requêtes). Nous cherchons à matérialiser une dispersion dans les résultats ramenés par une requête. Cette dispersion est un indice fort du potentiel de la requête à générer de l'ambiguïté, soit parce qu'une requête a une forme qui renvoie à des lexèmes identiques, soit parce qu'une requête se compose de différents aspects. Pour cela, nous utilisons le deuxième niveau de classification, les catégories thématiques. Au nombre de 6, elles ont été présentées en 5.2 page 80.

L'hypothèse est que les catégories, en nombre réduit et correspondant à un classement adapté pour les news, vont donner une approximation des domaines présents dans l'actualité et nous permettre d'observer la dispersion des résultats de la requête dans la base documentaire. Par cette démarche, nous cherchons à tester la capacité du processus de catégorisation à révéler une diversité d'interprétations possibles pour une requête donnée, et par conséquent, à déterminer si les catégories thématiques sont un outil de mise au jour de l'ambiguïté.

Nous allons tester cette hypothèse en deux temps. Dans un premier temps, nous allons déployer la catégorisation et vérifier les biais pouvant influer sur

la technique. Dans un deuxième temps, nous allons nous appuyer sur la confrontation de la catégorisation avec une méthode différente (Wikipédia) pour tester les catégories thématiques comme outil de mise au jour de l'ambiguïté.

6.1.2 Méthode

L'expérimentation consiste à catégoriser les requêtes pour pouvoir observer leur distribution sur plusieurs catégories thématiques. Les requêtes d'une période temporelle donnée sont projetées sur le corpus *2424beta* correspondant à la même période temporelle (par exemple Mai 2010). Si elles apparaissent dans un document, elles héritent de la catégorie thématique du document.

Les catégories sont pondérées selon la fréquence d'apparition. Nous effectuons également un filtrage des catégories attribuées : une catégorie est considérée seulement si elle représente plus de 10% des textes où les termes de la requête apparaissent. Le filtre permet de limiter l'apparition de catégories résiduelles et a été choisi de façon arbitraire.

Le corpus de requêtes est constitué à partir du corpus décrit dans le chapitre 5. Nous avons effectué une première sélection des requêtes les plus fréquentes, le corpus 2424reqFréquentes. Cette sélection a déjà été utilisée pour les premières observations présentées dans le chapitre 5 : 49 requêtes les plus fréquentes de chaque sous-partition du corpus et ayant une fréquence supérieure à 100 pour la période considérée. Le corpus utilisé contient 391 requêtes.

Une deuxième sélection est opérée, nous avons choisi de limiter l'étude aux requêtes mono-termes, qui sont plus affectées par l'ambiguïté (Sanderson, 2008). Ce filtrage est opéré par la projection des requêtes sur les documents, cela concerne 35% des 391 requêtes. Les requêtes contenant plusieurs mots et ne formant pas des termes sont exclues comme par exemple *boue hongrie*, alors qu'une requête comme *festival de cannes* va être conservée et catégorisée. Le corpus de requêtes catégorisées comporte donc 248 requêtes.

6.1.3 Premières observations

La répartition entre les requêtes mono-catégorielles et pluri-catégorielles est la suivante : 54% sont mono-catégorielles et 46% pluri-catégorielles. Nous observons que la répartition des requêtes qui donnent lieu à un classement thématique unique varie selon les corpus-tests (environ 67% pour le sous-corpus de décembre contre 25% dans le sous-corpus de mai) comme on peut le voir dans le tableau 6.1. Les résultats varient de manière très importante selon les sous-corpus. Les sources de cette variation ne sont pas connues.

Corpus	mai	juin	juillet	août	sept	oct	nov	déc
Mono-cat	25	48	62	55	70	46	50	67
Pluri-cat	75	51	37	44	29	53	50	32

TABLEAU 6.1 : Répartition entre requêtes mono-catégorisées et pluri-catégorisées selon les sous-corpus en %

Par exemple, *miss france* va être catégorisée exclusivement en CULTURES tout comme les requêtes *prince william* ou *audrey pulvar*. Ce sont des requêtes désignant des personnalités publiques qualifiées dans les médias de « people ». La requête *nicolas dupont-aignan* est catégorisée en ÉCONOMIE, domaine où cet homme politique s'exprime régulièrement. Ces requêtes contiennent des noms propres complets ce qui aide à l'identification, mais il y a également des requêtes mono-mot qui n'ont qu'une seule catégorie comme *vogica* en INTERNATIONAL (entreprise en faillite).

Les requêtes rattachées à plusieurs catégories peuvent être à la fois des EN comme la requête *obama* qui est catégorisée en ÉCONOMIE et INTERNATIONAL et des noms communs comme la requête *otages*. Cette requête est catégorisée en INTERNATIONAL et SOCIÉTÉ en juillet 2010. Les documents étiquetés en INTERNATIONAL contenant la requête *otages* regroupent plusieurs évènements impliquant des prises d'otages à l'étranger comme les otages des Farcs ou un rebondissement dans l'affaire des otages des JO de Munich en 1972. La catégorie SOCIÉTÉ est appliquée à des documents qui relatent les dernières informations sur les deux otages français détenus en Afghanistan, Hervé Ghesquière et Stéphane Taponier aujourd'hui libérés. L'exemple de la requête *royal* (en juin 2010) le montre également. Elle peut renvoyer à une EN ou un adjectif. La catégorisation fait ressortir deux aspects intéressants, d'une part une catégorisation en POLITIQUE et d'autre part une catégorisation en SPORT. La catégorisation en POLITIQUE de *royal* renvoie à *Ségolène Royal*, contrairement à la catégorisation en SPORT qui renvoie à l'adjectif *royal* présent dans un certain nombre de noms de stades sud-africains utilisés lors de la Coupe du monde de football, par exemple le stade Royal Bafokeng de Rustenburg.

Nous allons définir dans nos analyses des types élémentaires permettant de distinguer les requêtes composées d'EN (EN) et celles composées de noms communs (NC). Au sein des EN, nous avons opéré des distinctions : NPP (nom propre de personne), NPL (nom propre de lieu), EN_autres (principalement des noms d'organisations et d'entreprise).

Si l'on répartit les requêtes en types élémentaires (NC, EN), on constate que 60% des requêtes pluri-catégorisées sont des EN et 40 % sont composées de

NC. Cependant, les EN sont majoritaires dans notre corpus.

Les requêtes contenant des NC ont tendance à ouvrir vers plusieurs catégories comme on peut le voir sur la figure 6.2, par exemple, les requêtes *grèves* (ÉCONOMIE et SOCIÉTÉ), *ministre* (ÉCONOMIE, INTERNATIONAL, POLITIQUE et SOCIÉTÉ) ou encore *sport* (INTERNATIONAL, SOCIÉTÉ et SPORT). Alors que seulement 36 % de ces requêtes ne sont rattachées qu'à une seule catégorie.

Concernant les requêtes contenant des EN, nous observons une répartition différente puisque environ 60% des requêtes composées d'EN opèrent un rattachement unique et 40% sont rattachées à plusieurs catégories. Nous avons mesuré la répartition des rattachements simples et multiples pour deux types d'EN les plus importants en nombre, les NPP et les NPL. On voit dans la figure 6.3 que ce sont les requêtes mono-catégorielles qui sont majoritaires pour ces deux types d'EN. En effet, 63% des EN NPP (*johnny hallyday* (CULTURES)) et 56 % des EN NPL sont mono-catégorisées *(liban* (INTERNATIONAL)), et seulement 36% des NPP (*sarkozy* (INTERNATIONAL, SOCIÉTÉ et POLITIQUE)) et 43% de NPL sont rattachées à plusieurs catégories comme *corée* (INTERNATIONAL et SOCIÉTÉ). Ces résultats indiquent que les EN sont moins enclines à la pluri-catégorisation que les requêtes NC.

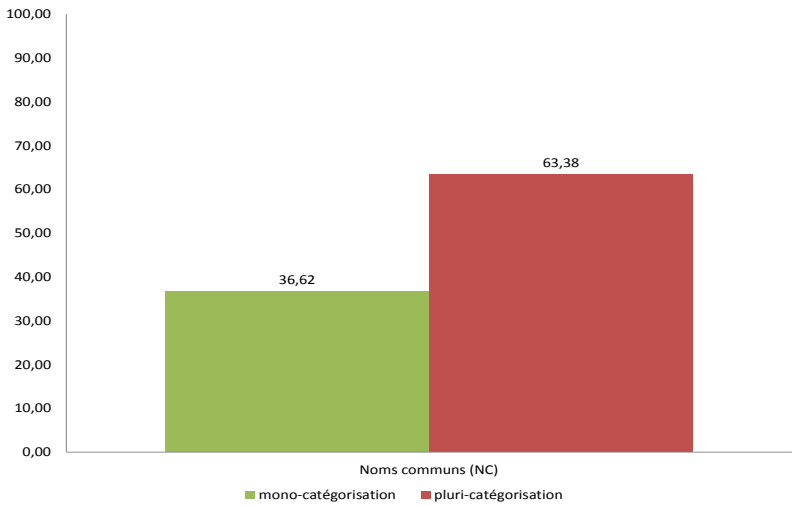


FIGURE 6.2 : Répartition des rattachements catégoriels des requêtes NC (en %)

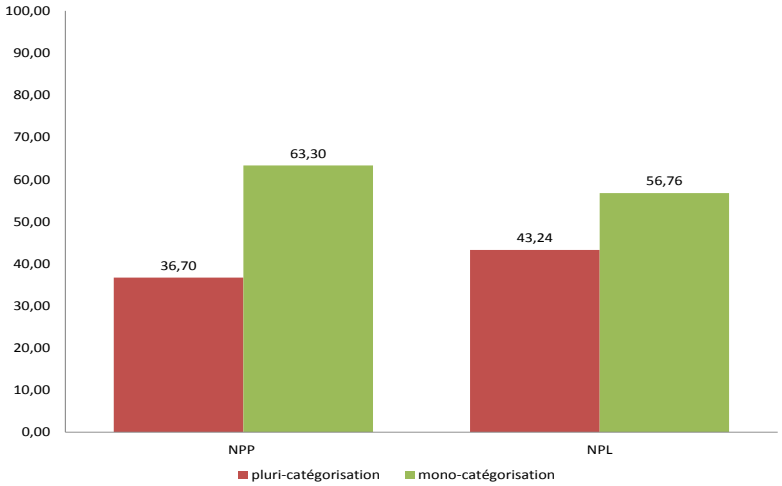


FIGURE 6.3 : Répartition des rattachements catégoriels des requêtes NPP et NPL (en %)

6.1.4 Examen des biais possibles de la catégorisation

Avant de progresser dans l'étude de l'apport de la catégorisation thématique, nous voulons tester les biais possibles de cette méthode. Nous avons identifié deux facteurs possibles pouvant faire varier la catégorisation. Le premier facteur est celui de la fréquence d'apparition des mots de la requête catégorisée dans les documents. La vérification va consister en un examen de la fréquence en fonction du nombre de catégories attribuées. Le deuxième facteur est celui de la distribution des catégories thématiques en elles-mêmes. Nous voulons observer si la répartition des requêtes selon les catégories est homogène.

6.1.4.1 La fréquence d'apparition de la requête dans les documents

La catégorisation s'appuie sur l'apparition des mots de la requête dans les documents étiquetés au préalable. Un biais possible est alors que la fréquence du mot ou du groupe de mots recherchés dans les documents influe sur le nombre de catégories proposées par la classification. Ainsi, un mot fréquent dans les documents serait susceptible d'apparaître dans des documents catégorisés différemment. Pour tester ce biais possible, nous avons représenté le nombre de catégories en fonction de la fréquence. Les résultats sont visibles dans la figure 6.4.

Les effets de fréquence sont limités dans cette représentation par l'application d'un logarithme en base de 10. La figure 6.4 montre que la fréquence augmente avec le nombre de catégories attribuées, mais que l'effet ne semble pas massif.

Nous notons que les individus rattachées à 5 catégories ou qui apparaissent un très grand nombre de fois dans les documents sont peu nombreux. Ainsi, seulement deux requêtes sont rattachées à 5 catégories, ce sont les requêtes *france* (novembre 2010) avec une fréquence de 2 380 et *nicolas sarkozy* (décembre 2010) avec une fréquence de 154. Les individus les plus fréquents sont :
– *ministre* (mai 2010), 4 catégories, fréquence de 3202 ;
– *ministre* (novembre 2010), 4 catégories, fréquence de 2744 ;
– *gouvernement* (juin 2010), 4 catégories, fréquence de 2547 ;
– *gouvernement* (mai 2010), 3 catégories, fréquence de 2736.

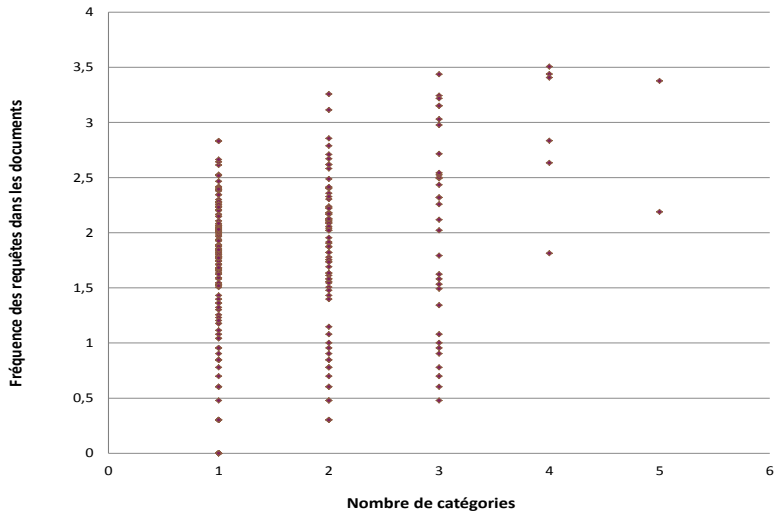


FIGURE 6.4 : Confrontation du logarithme de la fréquence des requêtes (base 10) par rapport au nombre de catégories rattachées à celles-ci.

Nous avons conforté les observations opérées sur la figure 6.4 par un calcul de coefficient de corrélation. Les résultats sont présentés dans le tableau 6.5. Deux calculs de coefficient ont été établis à partir du logarithme des fréquences :
– coefficient sur l'ensemble des requêtes étudiées,
– coefficient sur l'ensemble à l'exception des quatres requêtes les plus atypiques (*ministre* et *gouvernement*),

Le coefficient de corrélation est de 0,33 sur l'ensemble, ce qui est une valeur de corrélation faible. Cette valeur baisse encore lorsqu'on enlève les quatre individus atypiques décrits ci-dessus. Il n'y a donc pas de corrélation significative

Corrélation de Pearson	Valeur
Sur l'ensemble des requêtes	0,33
Sur l'ensemble - 4 requêtes	0,27

TABLEAU 6.5 : Coefficient de corrélation entre le logarithme des fréquences d'apparition des requêtes dans les documents et le nombre de catégories rattachées

entre la fréquence d'apparition des requêtes dans les documents et le nombre de catégorie attribuée. Par conséquent, la fréquence joue un rôle mais ce n'est pas le seul facteur explicatif.

6.1.4.2 Répartition des catégories thématiques

Les catégories thématiques sont établies au préalable et reposent sur une catégorisation expert. Le biais possible est que le filtre opéré par certaines catégories soit peu efficace, attirant la majorité des requêtes. Pour vérifier cela, nous allons analyser la répartition des catégories thématiques dans les deux types de cas, rattachement unique et multiple. La répartition des catégories ayant opéré des classements uniques est représentée dans la figure 6.6. La répartition des catégories étant retenues dans un classement multiple est visible dans la figure 6.7.

La répartition des catégories dans la figure 6.6 montre que les classements uniques s'opèrent majoritairement dans la catégorie INTERNATIONAL, à hauteur de 40%. La catégorie CULTURES occupe la deuxième place avec 20 % des classements mono-catégoriels. La catégorie SOCIÉTÉ est attribuée à 14% des requêtes mono-catégorisées. La catégorie SPORT est attribuée à 9% des requêtes mono-catégorisées, et les catégories ÉCONOMIE et POLITIQUE sont attribuées chacune à 6% des requêtes mono-catégories.

La répartition des catégories lors des classements multiples est plus équilibrée par rapport à la distribution des catégories dans les cas des classements uniques, comme nous pouvons le voir dans la figure 6.7. En effet, trois catégories thématiques sont particulièrement présentes : INTERNATIONAL (25 % des catégories attribuées), SOCIÉTÉ (25 % également) et ÉCONOMIE (19 %). La catégorie POLITIQUE est attribuée à 13% des requêtes pluri-catégorisées. La catégorie CULTURES est présente dans seulement 9% des classements multiples, répartition comparable pour la catégorie SPORT (8%).

L'observation de la répartition des catégories lors des classements uniques suggère donc que les catégories INTERNATIONAL et CULTURES sont trop vastes, englobant de nombreux sujets. Mais la répartition des classements multiples

nous donne une image plus nuancée de ces catégories. En effet, elles ne sont pas majoritaires dans les deux types de classement. Dans le cas des requêtes à classement multiple, nous voyons que les répartitions entre les catégories sont équilibrées. Par conséquent, les catégories thématiques ne semblent pas être un biais possible du processus de catégorisation.

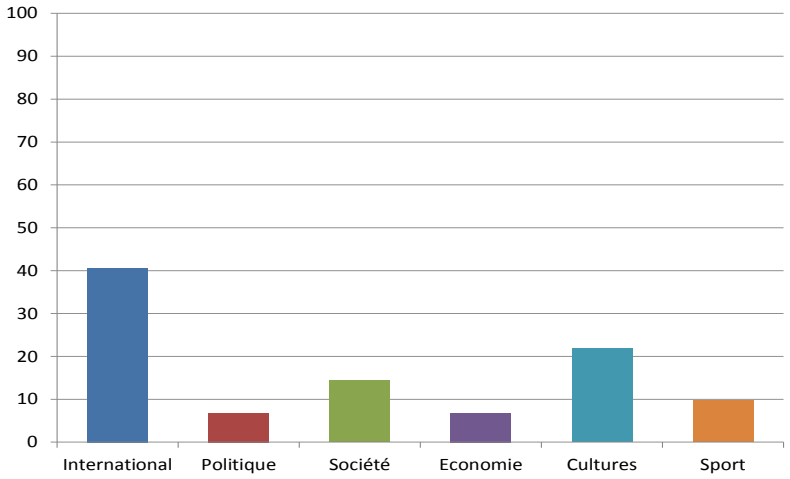


FIGURE 6.6 : Répartition des différentes catégories thématiques pour les requêtes mono-catégorielles (en %)

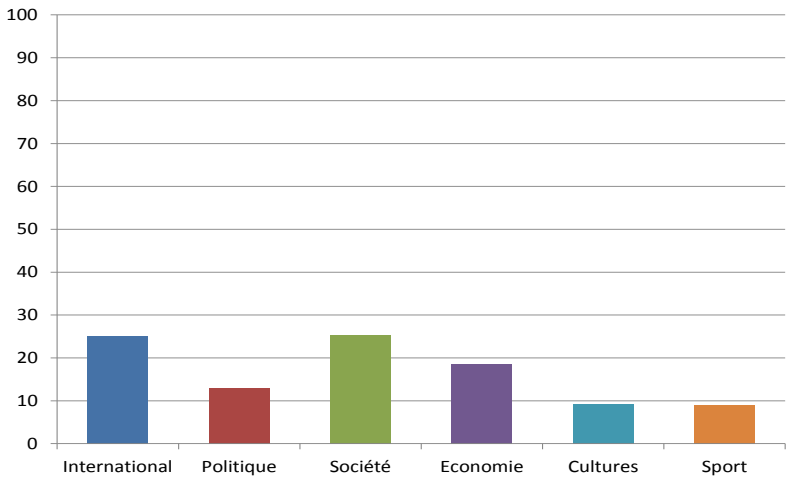


FIGURE 6.7 : Répartition des différentes catégories thématiques pour les requêtes pluri-catégorielles (en %)

6.1.5 Conclusion

Les résultats de la catégorisation montrent une tendance moins forte à la pluri-catégorisation pour les requêtes composées d'EN, malgré une prédominance des EN au niveau de l'ensemble du corpus. Les deux biais recherchés pouvant influencer la catégorisation s'avèrent ne pas avoir un impact significatif.

Le premier examen de ces résultats ne permet pas de donner du sens aux catégories thématiques en tant qu'outil d'observation. Nous allons donc évaluer notre méthode en la comparant à une autre. Pour cela, nous proposons d'utiliser le système de marquage de l'ambiguïté lexicale de Wikipédia (cf. 1.2.1). Nous savons que cette méthode ne va pas révéler le même type d'ambiguïté, et *a fortiori*, ne pas catégoriser les mêmes requêtes. Mais cette démarche va nous permettre de mieux observer le potentiel de la catégorisation thématique, et ses différences par rapport à une méthode plus traditionnelle.

6.2 Confronter deux sources de catégorisation : catégorisation thématique *versus* Wikipédia

Comme nous l'avons vu dans le chapitre 1, Wikipédia est la ressource la plus couramment utilisée pour recenser l'ambiguïté des requêtes. C'est une encyclopédie en ligne en libre accès. Sa construction repose sur certains principes comme le fait que chaque article décrit un seul « concept » et qu'il y a un seul article pour chaque concept. Une attention particulière est attachée à la création du titre d'un article, celui-ci doit être court et similaire aux termes d'un thésaurus conventionnel. Un autre principe de base est l'existence de pages de désambiguïsation. Elles servent à lister les possibles variations de sens d'un article (Medelyan *et al.*, 2009). Ces pages particulières ont été utilisées pour des travaux de désambiguïsation lexicale avec succès, améliorant de 30 à 40% les performances par rapport à une baseline Senseval [1] (Mihalcea, 2007). L'encyclopédie est également une ressource pour le TAL comme le montre le panorama de la question par Medelyan *et al.* (2009). En effet, c'est une ressource d'une certaine envergure, qui se développe grâce au travail collaboratif des internautes. Mais c'est également une ressource disponible en plusieurs langues, ce qui est une grande chance pour le TAL, où la majorité des ressources sont en anglais.

Par ailleurs, Wikipédia comporte de nombreuses entités nommées, en particulier grâce à une grande réactivité des « wikipédiens » face à l'actualité. Wikipédia était même vu en 2005 comme tendant à devenir un véritable « baromètre

1. Voir la section 2.1.2.

de sujets de société » (Klein, 2005). Cette tendance est toutefois à modérer en raison du ralentissement du développement de Wikipédia ces dernières années, en particulier pour le français [2]. Cette diversité autant sur les pages en elles-mêmes que sur la prise en compte de l'actualité nous paraît intéressante pour traiter nos corpus de requêtes.

C'est par la comparaison à Wikipédia que nous avons la possibilité de mieux étudier les différences de la catégorisation thématique. Cette ressource nous est apparu adaptée pour cette tâche car elle est à même de traiter les requêtes de notre corpus en comparaison aux autres ressources disponibles (voir le chapitre 1). Elle contient en effet de nombreuses entités nommées et elle est disponible en français.

Ainsi, nous avons réalisé tout d'abord une annotation qui permet de déterminer quelles requêtes peuvent être considérées comme étant ambiguës du point de vue de Wikipédia. La première étape de la confrontation entre les deux méthodes a été le calcul d'un coefficient de Kappa de Cohen, dans le but de confirmer que nous avions affaire à deux types de méthodes opérant différemment. Dans la deuxième étape, notre démarche a consisté à comparer l'annotation provenant de Wikipédia à la catégorisation thématique. C'est par ce procédé que nous pouvons étudier les différences qui existent et montrer la capacité de la catégorisation thématique à révéler la dispersion créée par une requête dans les résultats de recherche.

6.2.1 Annotation des requêtes avec Wikipédia

Nous avons confronté une partie du corpus *2424reqFréquentes* (97 requêtes, mai et décembre 2010) avec l'encyclopédie en ligne Wikipédia [3]. Deux aspects sont examinés :

– est-ce que la requête considérée a une page Wikipédia ?
– si la requête est une entrée de page d'encyclopédie, est-ce qu'elle renvoie vers une page d'homonymie ou de désambiguïsation [4] ?

Une page d'homonymie ou de désambiguïsation dans Wikipédia est simplement une page qui répertorie les différents sujets et articles partageant un même nom. Par exemple « éruption » renvoie vers une page qui recense un certain nombre d'éruptions :

– *une* ***éruption*** *volcanique en géologie ;*
– *une* ***éruption*** *cutanée ou rash en médecine ;*

2. http://stats.wikimedia.org/FR/TablesRecentTrends.htm
3. L'encyclopédie Wikipédia a été consultée pour cette expérience en janvier 2011
4. Lors de l'expérience, les deux types de marquages existaient. Aujourd'hui, seules les pages d'homonymie existent.

– *une **éruption** solaire, un phénomène très énergétique se produisant à la surface du Soleil ;*
– *un morceau de guitare électrique du groupe américain Van Halen, **Eruption** ;*
– *un groupe disco **Eruption**.*

L'annotation a été effectuée sur les 97 requêtes du corpus test, 73 requêtes ont pu être annotées grâce à Wikipédia. Les résultats donnent une proportion de 50 requêtes non ambiguës selon Wikipédia contre 23 qui font appel à la désambiguïsation, soit 31% de requêtes ambiguës.

6.2.2 Confrontation de Wikipédia à la catégorisation

La confrontation de la méthode Wikipédia à la méthode par catégorisation thématique n'a pu être opérée que sur 59 requêtes. Ces requêtes doivent en effet être annotées par Wikipédia et avoir été catégorisées thématiquement. Le tableau 6.8 rassemble les résultats issus de la confrontation des deux méthodes. Nous voyons d'emblée que les deux méthodes ne classifient pas de la même manière les requêtes.

comparaison	une catégorie	pluri-catégories	total
un sens Wikipédia	22	18	40
à désambiguïser Wikipédia	7	12	19
total	29	30	59

TABLEAU 6.8 : Comparaison Catégorisation et Wikipédia

Nous allons confirmer ces différences par un calcul du coefficient du Kappa de Cohen (Cohen, 1960). Cette mesure permet d'obtenir une mesure de l'accord entre les deux types de classification (catégorisation et Wikipédia). Le coefficient de Kappa (K) mesure l'accord inter-annotateurs sur une tâche de jugement en opérant une correction par rapport au hasard.

$$K = \frac{P(A)-P(E)}{1-P(E)}$$

où *P(A)* est la proportion d'accords observés et *P(E)* la proportion d'accord aléatoire.

Le coefficient obtenu est de 0,15 indique un accord très faible. Ce résultat confirme nos hypothèses. Elles ne repèrent pas les mêmes types de diversité portée par les requêtes.

Nous allons à présent comparer les deux méthodes que nous avons utilisées afin de comprendre la source de désaccord de ces deux méthodes. Pour cela,

nous examinons les différents cas d'accord et désaccord existant entre Wikipédia et la catégorisation thématique. L'étude des cas de désaccord va nous permettre de donner du sens à la pluri-catégorisation en observant les dispersions qu'elle révèle. Cette étude va consister à analyser les résultats de la catégorisation. L'analyse est basée sur le dépouillement de 10 contextes extraits du corpus pour chaque catégorie thématique rattachée à une requête étudiée. Nous cherchons à évaluer s'il est possible d'interpréter ces partitions.

Les requêtes sont analysées en fonction de leur type sémantique minimal comme en 6.1.3 : NPP (nom propre de personne), NPL (nom propre de lieu), EN_autres (principalement des noms d'organisations et d'entreprise) et NC (nom commun).

6.2.2.1 Cas d'accord n° 1 : un seul sens dans Wikipédia et une seule catégorie

Pour 22 requêtes, les deux méthodes sont en accord : Wikipédia les considère comme univoques et la catégorisation se limite à une thématique. Les requêtes concernées sont rassemblées dans le tableau 6.9. Dans ce tableau, nous avons indiqué la catégorisation de chacune des requêtes et le nombre de documents qui ont servi à la catégorisation. Ces requêtes sont en grande majorité des EN et nous observons une proportion importante de NPP.

Type de requête	Requêtes
NPP	*marine le pen* (POL (155)), *alain juppé* (INT (30)), *benoît xvi* (INT (54)), *jean-louis borloo* (POL (30)), *carla bruni* (CLT (12)), *audrey pulvar* (CLT (5)), *mahmoud ahmadinejad* (INT (4)), *nicolas dupont-aignan* (ÉCO (2)), *jean-luc mélenchon* (POL (1)), *tony parker* (CLT (1))
EN_autres	*wikileaks* (INT (169)), *ligue 1* (SPR (100)), *tempête xynthia* (SOC (45)), *miss france* (CLT (28)), *champions league* (SPR (1))
NPL	*afghanistan* (INT (142)), *pakistan* (INT (47)), *thaïlande* (SOC (1)), *cambodge* (INT (1))
NC	*agriculture* (ÉCO (7))

TABLEAU 6.9 : Requêtes univoques pour Wikipédia et mono-catégorisées

Parmi ces requêtes, nous remarquons un nombre assez important de requêtes dont la catégorisation repose sur très peu de documents. En effet, les requêtes

peuvent être très fréquentes et n'apparaître que peu de fois dans les documents. C'est un problème que ne résoud pas le filtrage opéré lors de la catégorisation.

6.2.2.2 Cas d'accord n° 2 : plusieurs sens dans Wikipédia et plusieurs catégories

Le deuxième cas d'accord concerne les requêtes qui sont considérées comme ambiguës par Wikipédia et rattachées à plusieurs catégories thématiques (12 requêtes). Ces requêtes sont présentées dans le tableau 6.10. On peut donc supposer que ces requêtes présentent une ambiguïté lexicale. Nous allons vérifier cela en examinant pourquoi Wikipédia les classe comme étant ambigues.

Les requêtes sont analysées en fonction de leur type sémantique minimal ce qui nous permet de faire des regroupements plus pratiques pour l'analyse.

Type de requête	Requêtes
NPL	*corée* (INT (71), SOC (9))
NPP	*bachelot* (ECO (8), SOC (15), SPR (6)), *bruni* (CLT (3), POL (8)), *obama* (ECO (36), INT (304)), *royal* (ECO (14), INT (19), POL (12)) , *sarkozy* (ECO (257), POL (542), SOC (109))
NC	*international* (CLT (46), ECO (173), INT (55)), *société* (ECO (114), INT (64), POL (45), SOC (84)), *transports* (ECO (109), SOC (78)), *éruption* (CLT (1), SOC (3)), *handball* (ECO (2), SOC (1), SPR (2))

TABLEAU 6.10 : Requêtes qui ont une page de désambiguïsation et qui sont pluri-catégorisées

La requête NPL La requête *corée* est la seule requête désignant un nom de lieu. Selon Wikipédia, la Corée peut prendre trois sens différents :

- la Corée ou péninsule coréenne, une région géographique et culturelle d'Extrême-Orient ;
- la Corée du Nord, officiellement appelée République populaire démocratique de Corée, pays localisé dans la partie septentrionnale de la péninsule coréenne ;
- la Corée du Sud, officiellement appelée République de Corée, pays occupant la partie méridionale de la péninsule coréenne.

L'ambiguïté portée par cette requête est de l'homonymie puisque la requête a une forme qui peut correspondre à deux spécifications.

Du point de vue de la catégorisation, cette requête a été rattachée à 2 catégories : INTERNATIONAL (71), SOCIÉTÉ (9). Le rattachement à la catégorie INTERNATIONAL révèle des documents où *corée* se spécifie en Corée du Sud ou en Corée du Nord. Le rattachement à la catégorie SOCIÉTÉ s'explique par la présence de la Corée du Sud dans un classement sur le niveau des élèves au niveau mondial comme on peut le voir dans l'exemple (1). La catégorisation opère donc une distinction différente de Wikipédia.

(1) Plus inquiétant : l'écart se creuse de plus en plus entre les bons et les mauvais élèves. (...) Comme lors de la précédente enquête de 2006, notre pays se situe dans la moyenne de l'OCDE pour les trois compétences étudiées (21ème sur 65 en compréhension de l'écrit, 22ème en mathématiques et 27ème en sciences), loin derrière la tête du classement 2009 composée de Shanghai, **Corée** du Sud et Finlande. [07/12/2010, RTL, SOCIÉTÉ]

Les requêtes NPP Cinq requêtes sont de type NPP comme on peut le voir dans le tableau 6.10. Wikipédia recense des ambiguïtés similaires pour les requêtes *bachelot*, *obama* et *sarkozy*. Ce sont tous des cas dans lesquels le patronyme est commun à différentes personnes. Par exemple, *obama* est le patronyme de Barack Obama, président des États-Unis, mais également celui de Michelle Obama, son épouse, et c'est également une ville située au Japon. Wikipédia signale donc une ambiguïté de type homonymique.

Les deux autres requêtes *bruni* et *royal* ont la particularité de pouvoir être une entité nommée ou un adjectif selon Wikipédia. En effet, *bruni* est à la fois le patronyme de Alberto Bruni Tedeschi, Carla Bruni-Sarkozy ou encore Dino Bruni qui est un coureur cycliste italien, et le participe passé du verbe *brunir*. Toujours selon Wikipédia, *royal* est un adjectif qualificatif et un patronyme. Il recense trois personnes portant ce patronyme : Ernie Royal, James Royal et Ségolène Royal.

Du point de vue de la catégorisation, les contextes d'apparition de ces requêtes donnent un éclairage différent. Par exemple, la requête *bachelot* est pluri-catégorisée en ÉCONOMIE (8), SOCIÉTÉ (15) et SPORT (6). Pour autant, elle ne désigne pas « Alexis Bachelot (1796-1837), missionnaire français à Hawaï » ou « François Bachelot (né en 1940), homme politique français, ancien membre du Front national » mais « Roselyne Bachelot-Narquin (née en 1946) femme politique française (UMP) ». Il n'y a pas de homonymie réelle dans les documents de notre corpus. Par contre, il existe bien une diversité de points de vue que traduisent les différentes catégories. Ces points de vue sont portés

par les nombreuses activités de Roselyne Bachelot, ministre de la Santé et des Sports en 2010.

Les requêtes NC Les cinq requêtes de type nom commun sont touchées selon Wikipédia par des ambiguïtés de type homonymique ou polysémique. Ainsi *international* peut être un adjectif ou un nom, *éruption* peut être un nom commun ou un nom propre.

Les requêtes *société, transports* et *handball* sont polysémiques selon Wikipédia qui liste plusieurs « sous-sens » comme par exemple pour la requête *société* :

1. Sens global de collectivité d'individus,
2. Formes ou aspects de sociétés,
3. Sens d'association,
4. Sens d'entreprise.

Toutefois, la polysémie repérée par Wikipédia semble peu compatible avec le contexte d'actualité. En effet, *handball* selon la page de désambiguïsation de Wikipédia est un « sport collectif où deux équipes s'affrontent avec un ballon sur un terrain rectangulaire. Sens courant de handball en Europe ». Cette page indique également l'existence du *handball gaélique*, « sport proche du squash ou de la pelote basque » et du *handball américain*, « sport dérivé du handball gaélique, sens courant du terme handball aux États-Unis ». Les sens les moins courants de *handball* ne sont en effet pas présents dans notre base documentaire.

Par conséquent, l'information apportée par la ressource encyclopédique est peu pertinente pour révéler l'ambiguïté potentiellement présente dans nos requêtes. Le diagnostic de l'homonymie de Wikipédia ne comprend pas les emplois effectifs dans le corpus.

6.2.2.3 Cas de désaccord n° 1 : un seul sens dans Wikipédia et plusieurs catégories

18 requêtes correspondent à cette situation. Elles sont présentées dans le tableau 6.11. Ces requêtes n'ont pas de page de désambiguïsation, mais elles ont été pluri-catégorisées. On pourrait supposer qu'elles ne sont tout simplement pas ambiguës. Mais le problème est que Wikipédia ne recense pas les ambiguïtés qui ne sont pas de l'ambiguïté lexicale, il se limite aux ambiguïtés de type homonymique. Des aspects comme les variations propres au contexte spécialisé ne figureront pas dans l'encyclopédie, aussi riche soit-elle. Il paraît donc indispensable d'essayer de comprendre pourquoi ces requêtes

ont été rattachées à plusieurs catégories. Pour cela, nous allons examiner les contextes d'apparition des requêtes dans les documents correspondant à la même période temporelle.

Notre but est d'essayer de comprendre cette catégorisation à la lumière des documents et de voir si celle-ci révèle une diversité d'emploi non repérée par Wikipédia. Pour cela, nous effectuons une analyse de résultats de la catégorisation grâce à Antconc (Anthony, 2011). Ce concordancier nous permet de dépouiller les contextes extraits et de déterminer la présence du/des mots recherchés dans un document pour une catégorie donnée dans un corpus donné. Pour limiter l'analyse lorsque les contextes sont trop nombreux, nous choisissons de limiter le nombre de contextes à 10, pour chaque catégorie rattachée, et ce de manière aléatoire.

Type de requête	Requêtes
NPL	*afghanistan*, *haïti* (2), *pakistan*, *irak*
NPP	*arnaud montebourg*, *brice hortefeux*, *françois baroin*, *jacques chirac*, *kate middleton*, *nicolas sarkozy*, *françois fillon*, *roman polanski*
EN_autres	*airbus*, *facebook*
NC	*gouvernement*, *ministre*, *salaires*

TABLEAU 6.11 : Requêtes qui n'ont pas de page de désambiguïsation mais qui sont pluri-catégorisées

Les requêtes NPL Les requêtes NPL sont au nombre de quatre, dont la requête *haïti* présente dans les deux sous-corpus étudiées (mai et décembre) :
– *afghanistan* (ÉCONOMIE (4), INTERNATIONAL (14), SOCIÉTÉ (13), mai 2010),
– *haïti* (CULTURES (1), INTERNATIONAL (1), SOCIÉTÉ (1), mai 2010),
– *haïti* (INTERNATIONAL (55), SOCIÉTÉ (34), décembre 2010),
– *pakistan* (ÉCONOMIE (9), INTERNATIONAL (22), POLITIQUE (14), SOCIÉTÉ (14), mai 2010).

Ainsi pour la requête *afghanistan*, les contextes reliés à la catégorie INTERNATIONAL font référence aux attaques de Talibans et au nombre de victimes liées au conflit, comme on peut le voir dans l'exemple (2). Les contextes reliés à la catégorie SOCIÉTÉ sont très homogènes et réfèrent tous aux otages français Hervé Ghesquière et Stéphane Taponier (exemple 3). Les contextes reliés à la catégorie ÉCONOMIE sont rares et semblent être la résultante d'une mauvaise thématisation.

(2) Un soldat de la force de l'Otan a été tué aujourd'hui par des insurgés dans le sud de l'**Afghanistan**, ce qui porte à 197 le nombre de militaires

étrangers décédés depuis le début 2010, a annoncé la force de l'Otan (Isaf)... [16/05/2010, LE FIGARO, INTERNATIONAL]

(3) 135 jours de détention pour les deux journalistes de France télévision en **Afghanistan**. Cela fait 135 jours qu'Hervé Ghesquière et Stéphane Taponier sont otages en **Afghanistan**. Samedi, Patrick de Carolis partira pour Kaboul [13/05/2010, FRANCE 2, SOCIÉTÉ]

En décembre 2010, la pluri-catégorisation de la requête *haïti* s'opère entre la catégorie INTERNATIONAL et SOCIÉTÉ. Lorsque le mot *haïti* est présent dans des documents étiquetés en SOCIÉTÉ, ces documents font référence aux différents évènements qui entouraient l'arrivée en France d'enfants haïtiens adoptés (exemple 4). Les contextes reliés à la catégorie INTERNATIONAL sont variés : les élections haïtiennes (exemple 5), les causes de l'épidémie de choléra (exemple 6).

(4) **Haïti**/adoptions : arrivée du 2d avion. Quatre-vingt-quatre enfants d'**Haïti** en cours d'adoption par des familles [24/12/2010, LEFIGARO, SOCIÉTÉ]

(5) **Haïti** à feu et à sang. Après le séisme et le choléra, **Haïti** affronte une nouvelle épreuve : la violente contestation des résultats du premier tour de l'élection [09/12/2010, JDD, INTERNATIONAL]

(6) WASHINGTON - Haïti : le bilan de l'épidémie de choléra dépasse les 2.500 morts WASHINGTON - L'épidémie de choléra en **Haïti** a fait plus de 2.500 morts depuis son apparition à la mi-octobre [19/12/2010, L'EXPRESS, INTERNATIONAL]

Les requêtes *pakistan* et *haïti (mai 2010)* sont plus difficiles à analyser du fait de leur faible nombre d'occurrences. La requête *pakistan* présente tout de même des contextes reliés à la catégorie SOCIÉTÉ homogènes, ils font référence à l'affaire Karachi, impliquant la France et le Pakistan. Les contextes reliés à la catégorie INTERNATIONAL regroupent des informations à propos d'attentats s'étant déroulés au Pakistan, dont certains impliquant les USA. On voit donc que ces noms propres de lieux peuvent exprimer différents points de vue et renvoyer à différents évènements susceptibles de constituer des blocs informationnels homogènes (cf. section 1.1.5).

Les requêtes NPP Les requêtes NPP sont au nombre de 8, dont 3 requêtes qui sont reliées à 3 catégories ou plus :

– *arnaud montebourg* (CULTURES (3), POLITIQUE (1), décembre 2010),
– *brice hortefeux* (POLITIQUE (36), SOCIÉTÉ (66), décembre 2010),
– *kate middleton* (CULTURES (3), ÉCONOMIE (1), décembre 2010),

– *françois fillon* (POLITIQUE (36), SOCIÉTÉ (54), décembre 2010),
– *roman polanski* (CULTURES (98), INTERNATIONAL (15), mai 2010).
– *françois baroin* (ÉCONOMIE (8), INTERNATIONAL (5), SOCIÉTÉ (3), décembre 2010),
– *jacques chirac* (CULTURES (6), ÉCONOMIE (1), POLITIQUE (1), décembre 2010),
– *nicolas sarkozy* (CULTURES (14), ÉCONOMIE (23), INTERNATIONAL (32), POLITIQUE (36), SOCIÉTÉ (25), décembre 2010).

L'analyse des contextes d'apparition des requêtes *arnaud montebourg, brice hortefeux, françois fillon* et *françois baroin* met en évidence différents points de vue ou rôles présents dans les documents à propos de ces personnes. Par exemple, les contextes d'apparition de la requête *arnaud montebourg* reliés à la catégorie POLITIQUE réfèrent aux ambitions présidentielles de cet homme politique alors que les contextes reliés à la catégorie CULTURES traitent spécifiquement de sa relation avec Audrey Pulvar, une journaliste. L'analyse des contextes d'apparition de la requête *françois baroin* montre que lorsque le document est étiqueté en ÉCONOMIE, le rôle de ministre du budget est souvent précisé comme dans l'exemple (7). Alors que dans les contextes reliés aux catégories INTERNATIONAL et SOCIÉTÉ, c'est le rôle de porte-parole du gouvernement, qui est mis en avant (exemple 8).

(7) Baroin : "Nous avons fait revenir 7 milliards de capitaux en France" **François Baroin**, ministre du Budget... [11/12/2010, JDD, ÉCONOMIE]

(8) Otages/Afghanistan : libération en cours **François Baroin**, porte-parole du gouvernement, a assuré aujourd'hui ... [21/12/2010, LE FIGARO, INTERNATIONAL]

L'analyse nous montre également que la catégorisation thématique peut présenter des imperfections. Certains choix de catégories assignées à un document méritent d'être revus comme pour la requête *kate middleton* et *jacques chirac* en décembre 2010. Par exemple, l'analyse des contextes montre que la catégorisation d'un document en ÉCONOMIE est une erreur. Seule la catégorie CULTURES subsiste, et les documents rattachés à cette catégorie offrent des contextes assez homogènes où Kate Middleton est identifiée comme la future épouse du prince William. Cet exemple montre également les limites de notre système de filtrage qui retient une classification basée sur un document lorsque les requêtes apparaissent dans peu de contexte.

Nous voyons à présent le cas de la requête *nicolas sarkozy* qui est rattachée à plusieurs catégories. Les contextes d'apparition de la requête *nicolas sarkozy* sont nombreux par rapport à ceux des requêtes étudiées ici. La catégorie SOCIÉTÉ révèle l'implication du « président » Nicolas Sarkozy dans des affaires de société, en l'occurrence l'affaire du Médiator comme on peut le voir dans

l'exemple (9). La catégorie INTERNATIONAL comprend des documents qui rendent compte de l'activité du président Sarkozy au niveau de la politique étrangère comme on peut le voir dans les exemples (10) et (11). La catégorie ÉCONOMIE comprend également des documents parlant du président Sarkozy (exemple 12). Les contextes issus des documents catégorisés en POLITIQUE ont beaucoup de points communs avec les documents rattachés aux catégories SOCIÉTÉ, ÉCONOMIE et INTERNATIONAL. Enfin, la catégorie CULTURES regroupe des contextes où Nicolas Sarkozy est avec son épouse Carla Bruni-Sarkozy, en particulier pour les fêtes de fin d'année (exemple 13).

(9) Mediator : Sarkozy a demandé "la transparence la plus totale". Le président **Nicolas Sarkozy** a demandé ... [22/12/2010, AFP, SOCIÉTÉ]

(10) INDE : Un allié incontournable. Le président Nicolas Sarkozy continue son voyage officiel en Inde. [06/12/2010, FRANCE 24, INTERNATIONAL]

(11) Nicolas Sarkozy : «Laurent Gbagbo doit laisser le pouvoir». La mise au point est intervenue lundi matin... [06/12/2010, LE PARISIEN, INTERNATIONAL]

(12) Sommet franco-allemand sur fond d'écart de croissance. Aux côtés de **Nicolas Sarkozy** et *François Fillon*, 8 ministres français participent à un sommet bilétaral à Fribourg. En 2010, Berlin a connu une croissance supérieure à 3,4 %, creusant l'écart avec la France. [10/12/2010, LA TRIBUNE, ÉCONOMIE]

(13) C'est déjà Noël à l'Elysée. Comme tous les ans, le président **Nicolas Sarkozy** et son épouse *Carla Bruni* ont accueilli des enfants à l'occasion du Sapin de Noël de l'Elysée. Cette année, les enfants victimes de le tempête Xynthia, en février dernier dans le Sud-Ouest du pays, ont reçu cet honneur, ainsi que d'autres enfants défavorisés. Au total, quelque 900 bambins sont repartis les bras chargés de cadeaux. [16/12/2010, PARIS MATCH, CULTURES]

L'analyse des contextes d'apparition de la requête montre que ces requêtes ne sont pas ambiguës au niveau linguistique. Leur référent est identifiable et il n'y a pas d'homonymie. Par contre, on remarque que ces requêtes peuvent avoir plusieurs « facettes », ce qui crée plusieurs interprétations possibles, liées à la présence d'évènements différents. La pluricatégorisation est révélatrice des types d'évènements auxquels réfèrent un terme. Wikipédia ne peut repérer ce phénomène. En effet, ce type de fonctionnement est difficile à imaginer au préalable, il est largement dû à des variations contextuelles. Nous rapprochons ces requêtes du type d'ambiguïté décrit dans le chapitre 1 en 1.2.2. Ce sont des requêtes « larges » (Song *et al.*, 2009) qui comportent plusieurs facettes.

Les requêtes EN_autres Les deux requêtes pluri-catégorisées que nous avons classées dans les EN_autres sont des noms d'entreprise :
– *airbus* (ÉCONOMIE (14), INTERNATIONAL (38), mai 2010),
– *facebook* (SOCIÉTÉ (78), CULTURES (5), mai 2010).
Prenons l'exemple de la requête *airbus*. Ce terme apparaît dans des documents catégorisés en ÉCONOMIE. Ceux-ci offrent des contextes qui permettent d'identifier l'entreprise Airbus, comme dans les exemples (14) et (15). Dans les documents catégorisés en INTERNATIONAL, le mot *airbus* ne réfère pas à l'entreprise mais à un avion ; *airbus* est alors modifié par un déterminant *un* ou *le*. Cet emploi est du à un crash d'avion à Tripoli, en l'occurrence un airbus A330 (exemple 16). Cette requête est donc ambiguë du point de vue référentiel et seul le contexte permet de la désambiguïser.

(14) France : Les traditionnels défilés syndicaux du 1er mai ont moins mobilisé que d'habitude. (...) Il y a eu un peu plus de 300 manifestations aujourd'hui, avec des usines en lutte, comme Unilever à Marseille, en grève pour les salaires, ou Freescale, Molex et **Airbus** à Toulouse. La CGT a compté 350.000 manifestants sur toute la France, à peine 200.000 selon la police ; [01/05/2010, FRANCE 3, ÉCONOMIE]

(15) Un euro en baisse, c'est bon pour les exportations, pas pour la consommation (...) C'est le cas des entreprises du secteur aéronautique et de défense, comme EADS, sa filiale **Airbus** et ses sous-traitants (Thales, Safran, Latécoère...), du luxe (LVMH) ou encore des producteurs de vins et spiritueux, qui se plaignent tous de manière récurrente de la force de l'euro. Pour Louis Gallois, le président d'EADS, (...) [04/05/2010, AFP, ÉCONOMIE]

(16) TRIPOLI - Libye : un enfant de huit ans survit au crash d'un avion qui a fait plus de 100 morts. TRIPOLI - Un garçon néerlandais de huit ans est le seul survivant du crash d'un avion, un **Airbus** A330 en provenance de Johannesburg, qui s'est écrasé mercredi à l'aéroport de Tripoli, faisant plus d'une centaine de morts, a-t-on appris de source aéroportuaire. [12/05/2010, L'EXPRESS, INTERNATIONAL]

La requête *facebook* est doublement catégorisée en SOCIÉTÉ et CULTURES. Les contextes issus des documents catégorisés en CULTURES sont moins nombreux (seulement 5) pour cette requête. Ces contextes peu nombreux concernent l'entreprise Facebook « nouveau n°1 américain » et le réseau social Facebook. L'analyse des contextes issus de documents catégorisés en SOCIÉTÉ, ne fait pas apparaître d'ambiguïté lexicale. Le mot *facebook* est toujours identifiable comme étant un réseau social (et non une entreprise) comme on peut le voir dans l'exemple (17). Nous observons lors du dépouillement de ces contextes

d'apparition que le mot *facebook* est associé au mot *apéro* pour désigner un type d'évènement qui s'est terminé de manière dramatique à Nantes (exemple 18). Cette requête peut donc être rapprochée des requêtes « larges » décrites ci-dessus.

(17) Trois salariés licenciés pour avoir critiqué leur hiérarchie sur **Facebook**. Trois salariés d'une entreprise de Boulogne-Billancourt (Hauts-de-Seine) ont été licenciés pour avoir dénigré leur hiérarchie dans une conversation privée sur le réseau social **Facebook**, a-t-on appris jeudi auprès de leur avocat. L'affaire, révélée jeudi par France Info, remonte à décembre 2008.... [20/05/2010, LES ECHOS, SOCIÉTÉ]

(18) Le décès lors d'un apéro **Facebook** entraîne l'interdiction de ces réunions. Le casse-tête des apéros **Facebook**... Hier, à Nantes, la soirée s'est mal terminée, un jeune homme de 21 ans s'est tué en tombant d'un pont. [14/05/2010, FRANCE 3, SOCIÉTÉ]

Les requêtes NC

– *gouvernement* (ÉCONOMIE (879), INTERNATIONAL (670), POLITIQUE (237), mai 2010),
– *ministre* (ÉCONOMIE (745), INTERNATIONAL (870), POLITIQUE (428), SOCIÉTÉ (304), mai 2010),
– *salaires* (ÉCONOMIE (97), INTERNATIONAL (31), mai 2010)

Pour les deux premières requêtes, la pluri-catégorisation est liée à un grand nombre de contextes d'apparition comme nous l'avions vu en 6.1.4.1. En effet, le mot *gouvernement* apparaît 2742 fois dans le corpus du mois de mai, *ministre* 3202 fois (contre 166 apparitions dans le corpus pour le mot *salaires*).

Prenons le cas de la requête *gouvernement* qui est catégorisée en INTERNATIONAL, ÉCONOMIE et POLITIQUE. Les contextes issus des documents en INTERNATIONAL montrent que la plupart du temps le mot *gouvernement* est suivi d'un adjectif qualificatif comme *thaïlandais* ou *anglais* précisant la nationalité du gouvernement en question. Le mot *gouvernement* a besoin d'être spécifié. Les contextes issus des documents en ÉCONOMIE sont assez similaires aux contextes étiquetés en INTERNATIONAL. *Gouvernement* est également spécifié par un adjectif. On note cependant que lorsque le mot *gouvernement* est employé sans qualificatif, il s'agit du gouvernement français dans les contextes étudiés. Enfin, dans les contextes étiquetés en POLITIQUE, on retrouve seulement un emploi de gouvernement sans qualificatif : « le gouvernement ». En l'occurrence, il s'agit du gouvernement français sur la question de la réforme des retraites. Le contexte est indispensable pour identifier le référent du mot

gouvernement. De même, c'est le contexte utilisateur qui est important, par défaut pour l'utilisateur, *gouvernement* signifiera peut être « gouvernement français », alors que dans la base il y a un grand choix. Le mot *gouvernement* n'est pas ambigu lexicalement mais il n'est pas autonome du point du vue référentiel.

La pluri-catégorisation met donc en évidence un autre type de mots ambigus : ce sont des termes non autonomes référentiellement. Cela se traduit par une multitude de référents candidats pour une requête donnée. Il existe alors une ambivalence par rapport au référent recherché par l'utilisateur. Ce type d'ambiguïté n'est pas recensée dans l'encyclopédie Wikipédia.

6.2.2.4 Cas désaccord n° 2 : plusieurs sens dans Wikipédia et une seule catégorie

Le second cas de désaccord apparaît lorsque Wikipédia détecte une ambiguïté lexicale, alors que le système de catégorisation ne donne qu'une catégorie thématique. Cette situation touche 7 requêtes. Elles sont visibles dans le tableau 6.12. Nous avons trié ces requêtes selon les types sémantiques présents : nom propre de lieux (NPL) et nom propre de personnes (NPP).

Type de requête	Requêtes
NPL	*bagdad, festival de cannes, israël* (2)
NPP	*berlusconi, prince william, johnny hallyday*

TABLEAU 6.12 : Requêtes qui ont une page de désambiguïsation et qui ne sont pas pluri-catégorisées

Pour les requêtes NPL, Wikipédia surestime le nombre de sens possibles pour le mot donné. En effet, Wikipédia se veut exhaustif et le but de l'encyclopédie en ligne est de collecter un maximum d'informations. Ainsi par exemple, la requête *festival de cannes* est catégorisée seulement en CULTURES mais elle a une page de désambiguïsation qui liste les sens suivant :

– Le Festival de Cannes est, parmi les festivals de cinéma, le plus médiatisé au monde
– Cannes Lions International Advertising Festival
– Festival international des jeux qui a lieu à Cannes
– Festival de Cannes de Scrabble francophone

Or, seul le premier sens de cette liste existe dans la base documentaire. Le cas est identique pour la requête *bagdad*, pour laquelle Wikipédia recense 8

toponymes homonymes et deux films. Dans notre contexte d'actualité, c'est seulement la ville de Bagdad en Irak qui est présente. La page d'homonymie d'*israël* sur Wikipédia est également très complète, détaillant les toponymes, l'histoire ou encore les religions liées à cette forme lexicale. Autant de sens qui ne sont pas utilisés dans notre contexte.

Les requêtes NPP sont également soumises au même effet. Wikipédia recense parfois plus de sens qu'il n'en existe dans notre contexte spécialisé : Johnny Hallyday peut aussi être un cascadeur selon l'encyclopédie or l'actualité ne connaît que le chanteur. Selon Wikipédia *prince william* peut faire référence à :
– William de Cambridge (1982-), un membre de la famille royale britannique,
– Prince William, une localité du comté de Carroll dans l'Indiana,
– Baie du Prince-William, en Alaska Navire,
– Prince William (voilier), un navire-école britannique.
Dans notre contexte, seule la première acception est présente, faisant l'actualité du mois de décembre 2010 à propos de son futur mariage. Le contexte d'actualité permet « d'éliminer » certaines ambiguïtés lexicales, ce qui montre l'intérêt d'une méthode qui tient compte du contexte.

Enfin, la requête *berlusconi* montre un autre aspect du décalage entre les deux méthodes. Dans Wikipédia, si les dispositifs signalant l'ambiguïté sont appelés à évoluer, cette évolution suit le rythme des créations de nouvelles pages et non celui de l'actualité. Or, la catégorisation suit l'actualité et le flux des documents. Par conséquent, la requête *berlusconi* serait considérée comme étant ambiguë mais mono-catégorisée en décembre, alors qu'un mois plus tôt, elle aurait été pluri-catégorisée.

6.2.2.5 Conclusion

La confrontration des deux méthodes nous permet à la fois de mieux comprendre la catégorisation thématique et l'utilisation d'une ressource. En effet, si nous avons vu que la pluri-catégorisation repère d'autres types d'ambiguïtés que l'ambiguïté lexicale, nous avons également constaté qu'une ressource encyclopédique seule n'était pas adaptée pour traiter des requêtes sur l'actualité. L'aspect changeant de l'actualité n'est pas à négliger. Cependant, l'observation des contextes montre des nuances qui ne sont pas révélées par la catégorisation. Cette méthode en effet opère un filtrage qui reste approximatif comme on a pu le voir par exemple pour la requête *facebook*.

La pluri-catégorisation s'avère être un outil intéressant pour révéler une diversité d'interprétation pour une requête donnée. L'ambiguïté ne se limite pas à l'ambiguïté lexicale, élément que nous mettions en évidence dans le chapitre

1. Après les analyses effectuées ci-dessus et les apports de l'état de l'art, nous pouvons dire qu'il existe deux types de requêtes ambiguës visibles dans le tableau 6.13 :
– celles dont on ne peut pas identifier le référent immédiatement (plusieurs candidats référents possibles) (1)
– celles dont on identifie le référent mais qui présentent des facettes (2)

Tout d'abord, l'ambiguïté référentielle est un type « englobant » qui recouvre à la fois l'ambiguïté lexicale (homonymie et polysémie), la métonymie (polysignifiance) et les termes non autonomes du point de vue référentiel (les noms communs principalement). Le deuxième type n'est pas de l'ambiguïté classique, nous le rapprochons de ce que Song *et al.* (2009) appelent les requêtes « larges ». Nous complétons leur définition, en décrivant les requêtes larges comme étant des requêtes dont le référent est identifiable mais fragmentable, matérialisant différents évènements dans l'actualité. C'est l'apport contextuel qui permet de les interpréter et de préciser l'information recherchée.

Ambiguïté référentielle (1)		Requêtes larges (2)
Ambiguïté lexicale	Termes non autonomes	Facettes
corée	*ministre gouvernement*	*nicolas sarkozy facebook*

TABLEAU 6.13 : Synthèse des types d'ambiguïté des requêtes rencontrés

6.3 Conclusion

Nous avons montré que la nature de l'ambiguïté repérée est bien différente selon les méthodes utilisées. Les exemples de décalage que nous avons examinés montrent que la catégorisation thématique capte mieux la réalité des emplois du corpus. En effet, une ressource construite dans un but pérenne comme peut l'être une encyclopédie n'est pas appropriée à la détection de l'ambiguïté des requêtes qui varie avec la base documentaire mais aussi avec l'actualité. Ce décalage temporel peut alors causer des faux positifs ou le repérage d'un type d'ambiguïté ne reflétant pas la situation.

Ces résultats vont dans le sens de Clough *et al.* (2009) qui ont montré qu'il n'y avait pas de corrélation entre les indicateurs qu'ils ont utilisés pour découvrir les requêtes à aspects multiples (en l'occurrence les clics et les reformulations de requêtes) et le nombre de sens pour un mot donné dans Wikipédia. L'ambiguïté détectée à l'aide de Wikipédia est centrée sur l'ambiguïté lexicale. Même

si ce type d'ambiguïté est présent dans les requêtes que nous avons pu étudier, ce n'est pas le seul type d'ambiguïté. Les problèmes d'ambiguïté référentielle ou de requêtes larges ne sont pas pris en compte.

Même si la catégorisation a certains défauts, elle reste une piste intéressante et sérieuse pour discriminer les requêtes ouvrant vers une diversité d'informations. Les types d'ambiguïté mis en évidence par l'analyse des résultats de la catégorisation méritent d'être testés à grande échelle afin d'évaluer la portée de chacun des types.

De plus, l'utilisation d'un moteur de recherche nous paraît nécessaire pour tester véritablement le dispositif dans un contexte de RI. La projection telle que nous l'avons réalisée n'est pas assez robuste pour aller plus loin que l'analyse et l'évaluation par une ressource. Nous proposons donc d'évaluer la catégorisation thématique dans un contexte applicatif en la confrontant aux utilisateurs.

Nous pensons que dans un deuxième temps, la catégorisation thématique pourrait être confortée par l'utilisation d'indices contextuels comme la reformulation des requêtes ou la temporalité, hypothèse confortée par les résultats de Clough *et al.* (2009).

Chapitre 7

Pertinence de la catégorisation thématique pour les utilisateurs

La catégorisation s'est révélée être un outil exploratoire intéressant pour la mise en lumière de la diversité portée par une requête. Cependant, nous avons également vu qu'il n'y avait pas forcément de correspondance entre catégorie et sens d'une requête. Ainsi nous souhaitons tester dans ce chapitre si la catégorisation thématique est un dispositif qui permet à l'utilisateur d'avoir connaissance de la diversité sémantique portée par une requête. Cette évaluation prend la forme de deux expériences impliquant des utilisateurs.

Pour cela nous articulons ce chapitre en deux volets qui comprennent chacun une expérience à part entière. La première expérience cherche à tester ce dispositif de catégorisation thématique, alors que la seconde se veut plus exploratoire cherchant à comprendre la tâche de catégorisation en elle-même. En effet, si la première expérience confronte des utilisateurs à un dispositif incluant des résultats de recherche catégorisés et à des requêtes ciblées, la seconde expérience place l'utilisateur comme acteur de ce dispositif en lui demandant de réaliser la classification des résultats. Enfin, ce travail d'évaluation a nécessité l'utilisation et le paramétrage d'un moteur de recherche. Ce volet technique a permis de mettre en place des conditions d'expérimentation satisfaisantes et adaptées au contexte industriel.

7.1 Expérience 1 : la catégorisation thématique face aux utilisateurs

Nous avons testé la catégorisation thématique comme grille d'observation de la diversité portée par les requêtes des utilisateurs de 2424actu dans le chapitre

6. La catégorisation s'est avérée une porte d'entrée sur cette diversité permettant de mettre au jour différents types d'ambiguïté et les variations d'emploi qui peuvent toucher une requête. On peut supposer que cette porte d'entrée soit également un moyen pour l'utilisateur d'accéder à une information diversifiée. Dans cet objectif, nous allons confronter des utilisateurs à des résultats de recherche catégorisés thématiquement dans des conditions expérimentales.

En confrontant les utilisateurs à la pluri-catégorisation, nous cherchons à voir s'ils arrivent à comprendre les regroupements de documents qui sont opérés par le filtrage basé sur leurs catégories. L'hypothèse étant que si les utilisateurs comprennent ces regroupements, cela veut peut être dire que la catégorisation est un moyen de rendre compte de la diversité portée par une requête.

Nous allons donc nous attacher à tester les requêtes pluri-catégorisées, qui ouvrent sur plusieurs catégories. Pour cela, nous proposons aux utilisateurs de trouver le « point commun » entre tous les documents regroupés sous une catégorie, en résultat à une requête pluri-catégorisée, et de le verbaliser en quelques mots lorsqu'il existe.

Si les utilisateurs arrivent à exprimer un point commun en évaluant les documents rassemblés dans une catégorie, on peut dire que le regroupement des documents sous une catégorie est compréhensible. Si les points communs exprimés par différents utilisateurs sont semblables, on pourra dire que le regroupement sous cette catégorie est pertinent, révélant la diversité portée par la requête.

7.1.1 Mise en place de l'expérimentation

La mise en place de l'expériementation a rencontré quelques difficultés. Nous avons dû trouver suffisamment de sujets qui correspondaient aux critères pour passer l'expérimentation, particulièrement en matière de connaissance de l'actualité. La tâche demandée a dû être limitée dans un temps court, les sujets ayant pris sur leur temps de travail pour passer l'expérience. Mais la principale difficulté que nous avons rencontrée est liée au contexte industriel. En effet, au moment où est réalisée l'expérience (février 2012), l'application 2424actu avait été arrêtée quelques mois plus tôt, tout comme notre collecte de données. Le choix est donc fait d'utiliser des informations ayant marqué l'actualité et accessibles (requêtes et index correspondant).

Nous avons également dû intégrer un moteur de recherche et créer une interface de post-traitement. Nous avons en effet choisi de mettre en place un moteur de recherche opérationnel pour générer des résultats et, d'autre part,

une interface qui présente des données à l'utilisateur et trie celles que le moteur fournit. Les tests ne se déroulent donc pas directement sur le moteur de recherche, l'utilisateur ne proposant pas de requêtes. Une phase de pré-tests a été réalisée, qui a permis de régler des questions à propos de l'affichage et de la tâche demandée aux utilisateurs.

Nous allons à présent décrire les éléments nécessaires à l'expérimentation. Nous débutons par les aspects concernant le choix des données utilisées dans l'expérimentation et les critères retenus pour sélectionner les utilisateurs. Nous exposons ensuite les aspects techniques (moteur de recherche, interfaces).

7.1.1.1 Les données d'expérimentation

La question du choix des données a été prépondérante dans la préparation de l'expérimentation. Les requêtes ont été choisies dans le corpus de requêtes 2424suite (cf. 5.2.2.2) parmi les requêtes populaires. Nous avons porté notre choix sur des requêtes populaires qui ont la capacité d'ouvrir sur plusieurs catégories, et qui renvoient à des informations largement connues. Ce second aspect doit aider à pallier au fait que l'actualité utilisée pour l'expérimentation avait déjà un an existence. Nous souhaitions également pouvoir tester les requêtes en fonction de leur type sémantique, et par conséquent avoir un panel de requêtes composées de noms de lieux, de noms de personnes et de noms communs.

Le nombre de requêtes a été décidé en fonction du nombre d'utilisateurs nécessaire pour donner une validité à l'expérimentation. Sur les conseils de Valérie Botherel, ergonome à Orange Labs, il a été décidé de prendre 9 requêtes (3 par catégories) et de les tester sur 20 utilisateurs. Les 9 requêtes correspondant à tous les critères énumérés ci-dessus (fréquence, actualité marquante, période temporelle, nature de la requête) sont présentées dans le tableau 7.1. Ce tableau rassemble l'identifiant de la requête, sa forme, sa fréquence dans son corpus de provenance, le corpus en question et le rang de la requête en terme de fréquence dans son corpus d'origine.

Le protocole de l'expérience intégre une requête « test » qui n'est pas présentée dans le tableau 7.1. Celle-ci sert de support afin d'expliquer à l'utilisateur la tâche qu'il doit effectuer. C'est la requête *mam*, acronyme de Michèle Alliot-Marie, qui a été choisie comme requête « test ». Elle a une fréquence de 994 en février 2011, et est la 12ème requête la plus fréquente du corpus de février. Elle correspond également à une période marquante pour Michèle Alliot-Marie, puisque l'ancienne ministre a connu une actualité agitée en février 2011 qui a conduit à sa démission du gouvernement Fillon.

Identifiant	Requêtes	Fréquence	Corpus	Rang fréquence corpus
1	*afghanistan*	1235	fév11	10
2	*wikileaks*	6797	janv11	2
3	*berlusconi*	531	fév11	16
4	*tunisie*	2211	janv11	6
5	*laetitia*	5041	fév11	3
6	*grève*	3109	fév11	10
7	*egypte*	4111	fév11	4
8	*médicaments*	245	fév11	27
9	*météo*	245	janv11	25

TABLEAU 7.1 : Requêtes retenues pour l'évaluation utilisateur

7.1.1.2 Les utilisateurs

Nous avons déterminé plusieurs exigences quant au recrutement des utilisateurs, sujets de l'expérience. Ils ont principalement été choisis sur la base du volontariat et de certaines compétences attendues. Les compétences attendues ont été consignées dans une fiche de renseignements que chaque sujet remplissait avant l'expérimentation. Les exigences minimales pour être sujet de l'expérimentation sont d'être un utilisateur averti des nouvelles technologies et de s'intéresser régulièrement à l'actualité. Ainsi nous avons demandé aux sujets quels étaient leurs usages du web et leurs fréquences d'utilisation. Concernant l'actualité, nous avons focalisé le questionnaire sur leur manière d'accéder à l'information et demandé à quelle fréquence ils se tiennent au courant de l'actualité.

Les sujets avaient la possibilité de choix multiples, excepté sur la question de leur intérêt vis-à-vis de l'actualité. Le questionnaire est disponible en Annexe B.1 (page 210).

Le questionnaire détaille plusieurs choix possibles correspondant à leur manière de s'informer, ces choix ne sont pas exclusifs. Nous proposons quatre choix possibles :

- journaux télévisés
- radio
- presse écrite
  - format papier
  - format numérique
- plate-forme d'actualité en ligne

Ces catégories correspondent aux catégories présentes dans l'application 2424-actu, à l'exception des formats papier. Nous avons distingué les deux formes de presse écrite, car la presse numérique entre dans la composition des plates-formes d'actualité. Sous la dénomination de « plate-forme d'actualité en ligne », nous désignons les agrégateurs comme Google Actualités et les portails d'actualités en ligne comme Yahoo News ou le portail Orange.

7.1.1.3 Le moteur de recherche

Le choix s'est porté sur le moteur *open source* Terrier [1] (Ounis *et al.*, 2007), utilisé pour les campagnes d'évaluation TREC.

Nous avons utilisé la version Terrier - Terabyte Retriever version 3.0. Nous avons pratiqué une indexation avec élimination des mots vides, modèle Okapi BM25 [2] avec paramètres par défaut, reflétant la configuration optimale du système dans la majorité des situations. Pour cela, nous avons transformé le corpus de documents dans un format accepté par le moteur de recherche. C'est un format XML semblable au format des documents dans TREC comme on peut le voir dans la figure 7.2. Nous n'avons pas utilisé le stemmer proposé par Terrier, car il n'est pas adapté pour le français. Deux index ont été créés pour l'expérimentation, ils ont été constitués à partir du corpus *2424suite* [3].

Terrier nous permet donc d'obtenir les résultats à une requête et d'avoir un score de similarité entre chaque document retourné et la requête. Ces résultats nous servent à alimenter l'interface utilisée pour l'expérience.

```
<DOC>
<IDENT>5025150</IDENT>
<DATE>2011-01-18</DATE>
<SOURCE>SPORT24</SOURCE>
<THEME>spr</THEME>
<TITLE>Football - Vieira : «Materazzi, le plus con»</TITLE>
<TEXT>Patrick Vieira a évoqué son ancien partenaire à l'Inter Marco Materazzi
dans des termes fleuris. </TEXT>
</DOC>
```

FIGURE 7.2 : Exemple de document au format XML

1. http://terrier.org/
2. cf. chapitre 43 (3.1.2)
3. voir le tableau 5.10 (chapitre 77)

7.1.1.4 L'interface de recherche

L'interface développée effectue un post-traitement des résultats de recherche (figure 7.3). Elle filtre et affiche les résultats à une requête donnée, en fonction de critères précis. Elle a été conçue en php et assure trois tâches dans l'expérimentation :

– filtrer les résultats en provenance de Terrier ;
– afficher les résultats à l'utilisateur ;
– récupérer et stocker les informations ajoutées par l'utilisateur pendant l'expérience.

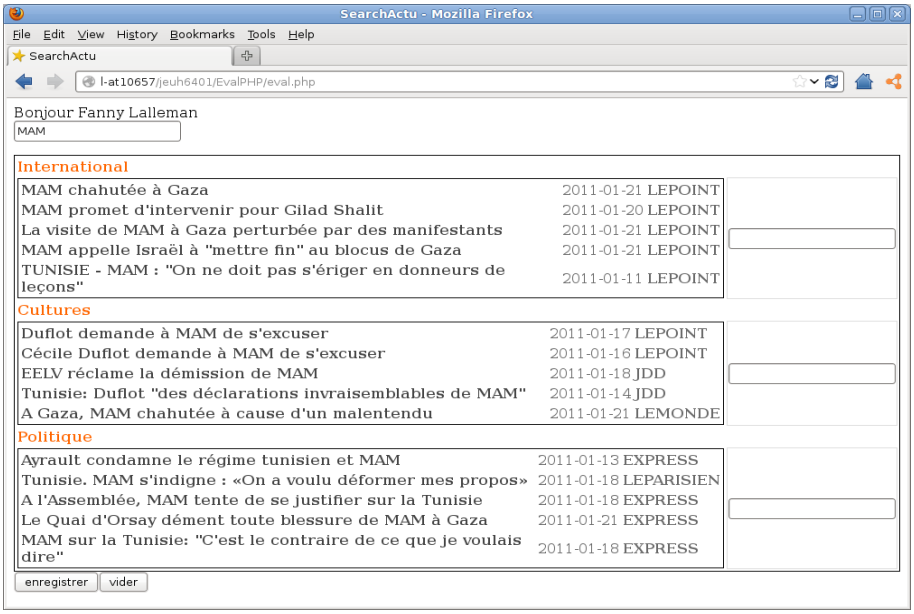


FIGURE 7.3 : Interface de tests pour l'expérience 1

Le filtrage est opéré grâce à la combinaison de 4 heuristiques, qui permettent d'opérer le tri des résultats via les catégories thématiques. En résumé, voici les quatre étapes suivies par le programme de post-traitement des résultats :

1. Filtrage par catégorie thématique : le filtrage repose, a minima, sur un premier filtrage par catégorie thématique. Les documents retournés par Terrier sont alors classés par catégorie thématique.
2. Filtrage par similarité dans chaque catégorie : tous les documents retournés par le moteur qui sont catégorisés par exemple en INTERNATIONAL sont alors réordonnés selon le degré de similarité avec la requête. Cela permet de faire remonter les documents les plus pertinents pour la requête donnée.

3. Choix des 4 meilleurs documents par catégories : pour des raisons d'ergonomie, il a été choisi de garder pour l'affichage les 4 meilleurs documents par catégorie thématique. L'affichage des catégories est aussi restreint par le nombre de documents retournés pour une catégorie. Si moins de 4 documents sont rattachés à une catégorie pour une requête, elle est jugée peu représentative et écartée.
4. Tri par longueur du document à afficher : le dernier tri concerne la longueur du document à afficher est effectué, les documents plus longs remontent dans le classement.

Nous avons choisi, après les pré-tests, de ne pas afficher le contenu du document directement dans l'interface, mais de laisser un accès au contenu via le titre. Les utilisateurs peuvent néanmoins accéder au document en passant la souris sur le titre.

7.1.2 Déroulement de l'expérience 1

L'expérience s'est déroulée du 27 février au 2 mars 2012, sur le site de Orange Labs à Lannion. L'ensemble des sujets recrutés pour l'expérience sont des collaborateurs de Orange Labs sur la base du volontariat. Enfin de décrire le déroulement de l'expérience, nous allons présenter le protocole de l'expérience, puis effectuer un focus sur les sujets de l'expérimentation. Cette description se termine par une revue des problèmes et difficultés rencontrés.

7.1.2.1 Le protocole de l'expérience

Le scénario de l'expérience s'articule autour de cinq étapes successives. Il débute avec l'accueil du sujet et se termine avec son départ, après un « débriefing ». Aucune limite de temps n'est imposée aux sujets. Voici une vue synthétique du scénario :

1. Remplissage du questionnaire par le sujet
2. Présentation de l'expérience au sujet
3. Consigne et exemple de la tâche avec la requête *mam*
4. Passation de l'expérience
5. Expérience terminée - Entretien de « débriefing »

Étape 1 Le sujet est accueilli, nous vérifions son niveau d'information sur l'actualité. Il remplit le questionnaire de renseignement.

Étape 2 C'est le moment le plus important de l'expérience, il doit permettre de mettre le sujet en condition pour effectuer la tâche. Le sujet est informé de la durée approximative de l'expérience et qu'elle sera suivie d'un court entretien. Il est ensuite informé du cadre de l'expérience. Celle-ci reproduit une situation de recherche d'information réelle dans un site regroupant l'actualité en temps réel. Par conséquent, un utilisateur qui interroge ce moteur d'actualité plonge dans l'actualité sous toutes ses formes (journaux télévisés sous la forme de vidéo, radio, dépêches AFP, articles de presse en ligne).

L'utilisateur est aussi averti avant de passer à l'étape 3, que les requêtes utilisées pour l'expérience ont été formulées par des personnes qui ont consulté le moteur du site dédié à l'actualité en question. Les résultats sont des sources d'actualité, retournées en réponse à une requête, disponibles à la même période.

Étape 3 Cette étape permet de décrire au sujet l'interface qu'il a face à lui. Il est informé que cette interface comprend deux types d'information :
- une requête, c'est-à-dire un mot-clé fréquemment formulé par des utilisateurs d'un site d'actualités.
- des titres de document classés par thématique, ces titres forment des « regroupements ».

Le sujet est également averti qu'il pouvait arriver que deux titres de documents soient identiques, simplement parce que deux journaux ont proposé des titres identiques, ou bien qu'un journal a corrigé son article et l'a republié.

Lorsque la consigne est expliquée à l'utilisateur il a devant lui la requête *mam* et les résultats retournés pour cette requête (cf. figure 7.3). On lui demande de regarder attentivement les titres de documents présentés en résultat à la requête proposée dans la partie haute de l'interface. Voici la consigne proposée au sujet :

> *Nous vous demandons d'indiquer, en quelques mots, comment vous interprétez chaque regroupement de news. Vous avez une zone prévue à cet effet à droite. Ces quelques mots doivent permettre d'identifier le point commun qui existe entre tout ou partie des documents présents dans le regroupement. Vous pouvez laisser la zone vide si vous ne réussissez pas à trouver de point commun à un ensemble suffisant de documents.*
>
> *Ces mots doivent permettre d'identifier le point commun que partagent selon vous un maximum de titres du « regroupement », même si ce maximum est « faible ».*
>
> *Pour passer à la page suivante, vous devez cliquer sur le bouton « enregistrer ».*

Ainsi le sujet doit identifier le « point commun » entre les documents présents dans le regroupement, et s'il voit plusieurs tendances se dégager, il lui est conseillé de choisir la tendance la plus importante. Enfin, s'il ne trouve pas de point commun, il doit laisser la zone de commentaire vide. Le sujet essaie alors de remplir les zones de commentaire prévues pour la requête *mam* et peut poser autant de questions qu'il est nécessaire.

Étape 4 Pendant cette étape, le sujet est laissé seul face à l'interface sur un ordinateur individuel. Mais nous restons à sa disposition dans la pièce s'il a une question. Il a donc 9 pages à évaluer selon la consigne portant chacune sur une des requêtes du panel choisi.

Étape 5 Une fois que le sujet a terminé son évaluation des requêtes, il doit répondre à quatre questions (étape 5) :
– *Avez-vous trouvé la tâche difficile ?*
– *Qu'avez-vous pensé des catégories proposées ?*
– *Est-ce que vous vous êtes servi de vos connaissances pour le test ?*
– *Avez-vous des questions sur l'expérience ou son but ?*
Ces questions doivent nous servir à évaluer si le sujet a compris ce qu'il lui était demandé de faire pendant l'expérience. Elles nous permettent également d'avoir un retour qualitatif sur notre expérience et de donner la possibilité au sujet de poser à son tour des questions sur notre travail.

7.1.2.2 Les sujets de l'expérience

Les pré-tests réalisés en amont de l'expérience en elle-même ont été effectués avec deux sujets. Ces personnes étaient des volontaires. Elles nous ont permis de préciser certains points comme la consigne de l'expérience, mais aussi les usages en terme d'information, et particulièrement la fréquence (une fréquence d'information quotidienne est recommandée).

L'expérience a été réalisée par 20 sujets (6 femmes/14 hommes). Ils ont été recrutés via un appel au volontariat diffusé sur l'ensemble du site de Lannion d'Orange Labs. Parmi les 20 sujets, 19 déclarent s'informer tous les jours et 10 d'entre eux s'informent même plusieurs fois par jour. Un seul sujet a déclaré s'informer 2 à 3 fois par semaine.

Les sujets sont a priori des grands « consommateurs » d'actualités car en plus d'une pratique quotidienne de l'information, le questionnaire met en évidence une combinaison de plusieurs moyens d'information. En effet, 95% des sujets

déclarent utiliser au moins 2 sources pour s'informer. En moyenne, ils utiliseraient 3,05 sources d'information de façon quotidienne, les trois sources les plus utilisées étant : les plates-formes d'actualité en ligne, la radio et les journaux télévisés (figure 7.4). La radio s'avère être le moyen le plus utilisé par les sujets : 17 sujets déclarent écouter la radio pour s'informer. On remarque également dans la figure 7.4 que la presse écrite papier et la presse écrite numérique sont utilisées de manière comparable par les sujets pour s'informer.

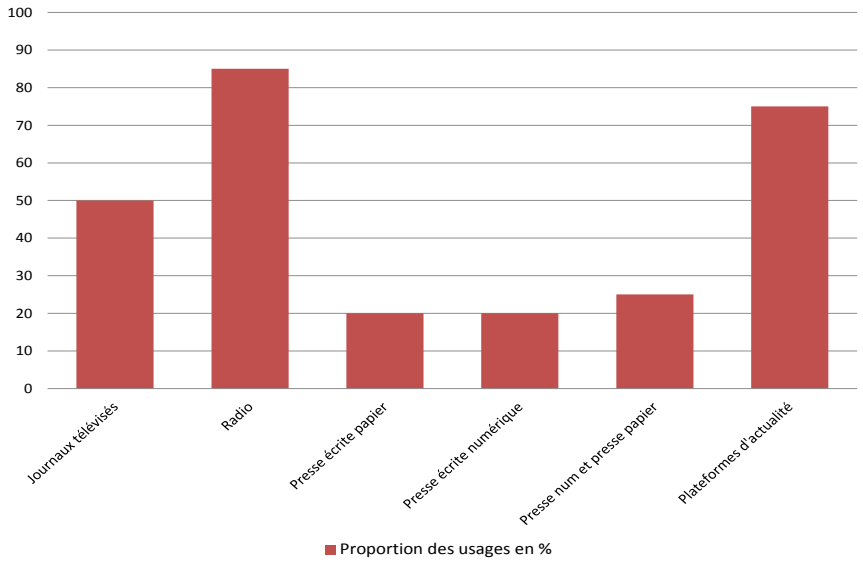


FIGURE 7.4 : Usages en matière d'accès à l'information des sujets de l'expérience 1

7.1.3 Résultats

En moyenne, les sujets ont mis 24,5 minutes pour effectuer l'expérience c'est-à-dire pour évaluer les 9 requêtes du panel, l'évaluation de la requête-test étant effectuée avec mon aide. Le sujet le plus rapide a mis seulement 14 minutes pour faire l'expérience, contre 45 minutes pour le sujet le moins rapide. Nous rappelons qu'aucune limite de temps n'était imposée pour accomplir l'expérience.

Nous distinguons plusieurs types de réponses parmi les réponses données par les 20 sujets :

– les réponses non vides
– les réponses vides

– les réponses identiques à la requête
Une réponse se matérialise sous la forme d'un « label ». Ce label est censé exprimer le point commun existant entre les documents retournés par le moteur et rattachés à une même catégorie thématique. Les réponses identiques à la requête sont considérées comme des réponses vides. En effet, si le sujet n'a pas donné de label autre que la requête en elle-même, c'est qu'il n'a pas trouvé de point commun entre les documents proposés. Les réponses dites « exploitables » sont donc les réponses non vides auxquelles nous avons soustrait les réponses identiques à la requête.

On totalise 534 réponses exploitables pour les 9 requêtes du panel sur 720 réponses effectives, ce qui représente 74,1% de réponses exploitables. L'ensemble des résultats requête par requête est disponible dans l'Annexe B.2.1 (page 211).

Les réponses vides ou identiques à la requête Nous notons de nombreuses reprises de la requête évaluée dans les labels, signe que le sujet n'a pas été en mesure de trouver un autre point commun et qu'il n'a pas identifié de diversité. Cependant, les réponses vides ou identiques à la requête doivent être également considérées parce qu'elles sont révélatrices d'une incompréhension ou d'un désaccord avec la catégorisation proposée.

Par exemple, dans le cas de la requête *laetitia*, deux catégories récoltent peu de réponses comme on peut le voir dans le tableau 7.5. Les résultats de cette requête sont répartis en 4 catégories : SOCIÉTÉ, POLITIQUE, CULTURES et ÉCONOMIE. Il s'avère que la catégorie SOCIÉTÉ regroupe des documents sur le meurtre d'une jeune femme prénommée Lætitia. Les médias n'ont communiqué que son prénom car elle était mineure. La catégorie POLITIQUE contient des documents qui parlent de la révolte des magistrats après que Nicolas Sarkozy, alors président de la République, les a rendu responsable du meurtre de la jeune femme. Ces deux catégories montrent les deux versants principaux de « l'affaire Lætitia ». Les deux autres catégories (CULTURES et ÉCONOMIE) rassemblent des actualités à propos d'autres personnes prénommées Lætitia dont particulièrement une journaliste qui anime quotidiennement une émission sur RTL, émissions qui sont accessibles depuis l'agrégateur 2424actu.

Les sujets ont évalué les deux premières catégories en soulignant l'intérêt de séparer les deux aspects de l'affaire Lætitia, les deux autres catégories leur ont paru totalement extérieures à la requête ou renvoyant des informations inconnues du sujet, et ils ont décidé de ne pas les annoter. Face à ce cas typique d'homonymie, ils semblent avoir choisi le sens « majoritaire » dans les résultats et rejeté les formes homonymiques.

Un phénomène identique s'est produit pour la requête *météo* (tableau B.10, Annexe B.2.1). Trois catégories se partageaient les résultats à cette requête : INTERNATIONAL, SOCIÉTÉ ET CULTURES. La catégorie INTERNATIONAL regroupe des documents parlant de la météo dans le monde et particulièrement dans les DOM-TOM. La catégorie SOCIÉTÉ contient des documents à propos des intempéries en France. La catégorie CULTURES contient principalement des documents qui parlent de la sortie d'un film sur la vie d'une Miss Météo. Les sujets de l'expérience n'ont pas compris pourquoi ces documents étaient présents.

leatitia	Réponses exploitables
POLITIQUE	18 / 20
SOCIÉTÉ	20 / 20
CULTURES	4 / 20
ECONOMIE	0 / 20
total	42 / 80

TABLEAU 7.5 : Résultats requête *leatitia*

Les réponses non vides La plupart des requêtes totalisent un taux de réponses très élevé pour l'ensemble des catégories que les sujets devaient labelliser. Parmi ces requêtes, nous retrouvons *wikileaks* (90% de réponses exploitables) ou encore *afghanistan* (95% de réponses exploitables). Par exemple, la requête *afghanistan* qui compte trois catégories, a obtenu des réponses pour l'ensemble de ces catégories (tableau 7.7). Les sujets ont donc réussi à trouver le point commun existant entre les différents documents organisés en catégorie thématique.

Dans le tableau 7.6, nous pouvons voir un exemple des réponses fournies par les sujets pour cette requête. Nous avons pris pour exemple l'ensemble des labels produits par les sujets pour la catégorie CULTURES. On remarque que les labels proposés par les sujets sont plutôt homogènes, plusieurs mots ressortent comme *otages, français* ou *journalistes*. Les documents retournés sous cette catégorie thématique parlent du soutien aux journalistes Hervé Ghesquière et Stéphane Taponier, journalistes enlevés en Afghanistan et retenus en otages durant de nombreux mois.

Nous constatons également que certains sujets ont pris des libertés avec la consigne, particulièrement sur la requête *berlusconi* ou la requête *égypte*. Par exemple, un sujet a donné pour label *Les vieilles habitudes ont une fin* aux documents de la catégorie POLITIQUE. Les documents concernaient la démission de

Hosni Moubarak. Le problème soulevé par ce type de label, c'est qu'ils ne permettent pas une analyse de manière quantitative, même si nous comprenons que le sujet a bien identifié le point commun présent dans les documents.

Labels proposés par les sujets			
1	otages français afghanistan	11	les plus long otages modernes français
2	durée de prise en otage de deux français en Afghanistan	12	Investissements médiatiques sur le sort des otages français
3	otages en Afghanistan	13	journalistes enlevés
4	herve ghesquiere stephane taponier otages afghanistan	14	otages afghanistan
5	reporters otages français en afganistan	15	otages français dans le monde
6	soutiens aux 2 journalistes otages	16	otages afghanistan durée
7	otages français afghanistan	17	reporter otages
8	Situation des otages français en Afghanistan	18	otages français en Afghanistan
9	afghanistan otages	19	statistiques sur enlevement des journalistes
10	Otages français en Afghanistan	20	otage - durée captivité

TABLEAU 7.6 : Labels proposés par l'ensemble des sujets pour la catégorie CULTURES de la requête *afghanistan*

afghanistan	Réponses exploitables
INTERNATIONAL	20 / 20
SOCIÉTÉ	17 / 20
CULTURES	20 / 20
total	57 / 60

TABLEAU 7.7 : Résultats requête *afghanistan*

7.1.4 Évaluation des résultats

Nous avons confronté les sujets de l'expérience à la pluri-catégorisation, en cherchant à savoir si les regroupements de documents qui sont opérés par

le filtrage des catégories thématiques font sens pour eux. Nous rappelons ici que notre hypothèse est que si les sujets comprennent ces regroupements, cela signifie que la catégorisation est un moyen de rendre compte de la diversité portée par une requête.

L'évaluation des résultats de l'expérience est articulée en deux volets : l'évaluation de la similarité des labels produits par les sujets (recouvrement inter-sujet) et l'évaluation des labels produits pour chaque groupement thématique pour une requête donnée.

7.1.4.1 Évaluation de la similarité des labels produits par les sujets

La principale difficulté est de pouvoir évaluer la similarité des labels produits par les sujets de l'expérience. Le recoupement des labels est un deuxième indice pour valider l'hypothèse selon laquelle la pluri-catégorisation des résultats est un dispositif compréhensible pour l'utilisateur. Un fort recoupement indique alors un accord certain entre les différents sujets, ce recoupement se traduisant par des labels indentiques ou ayant de nombreux termes en commun. Il est donc nécessaire d'évaluer pour chaque catégorie le recoupement des labels produits. La question est alors de savoir comment mesurer le recoupement qui existe potentiellement entre les labels d'une même catégorie.

Nous proposons de mesurer ce recoupement en regardant si chaque terme d'un label apparaît dans les autres labels de la même catégorie. Ceci permet alors de comparer les labels produits pour une même catégorie et une même requête. Une démarche manuelle s'impose car les labels comportent un grand nombre de fautes et variantes orthographiques, elle nous permet également d'effectuer une évaluation qualitative des labels.

Pour 9 requêtes testées, 36 groupements thématiques ont été évalués par les 20 sujets.

Nous avons choisi de ne calculer le recouvrement que pour les mots pleins. Les prépositions et les déterminants ne sont donc pas comptabilisés. Cela permet de se concentrer sur le sens porté par le label et de limiter les particularités linguistiques de chaque sujet. Nous considérons comme similaires deux mots écrits avec ou sans accent. Nous distinguons par contre les formes au singulier et au pluriel [4].

4. Nous avons opéré cette distinction car les documents présentés à l'intérieur d'un groupement thématique pouvaient parler d' « un soldat » ou « des soldats » et que cette distinction a été reprise par les sujets dans les labels.

Calcul du recouvrement entre labels inter-sujets intra-catégorie Le recouvrement entre labels est donc calculé en attribuant un score à chaque label, dont la formule est :

$$Score_i = \frac{N_i \cap L_n}{N_i}$$

N_i : ensemble des mots du label$_i$

L_n : ensemble des mots des autres labels appartenant à la même catégorie que celle du label$_i$

Ce score correspond au rapport entre le nombre de mots du label (N_i) qui sont présents dans les autres labels de la même catégorie ($N_i \bigcap L_n$) et le nombre de mots du label (N_i). Ce rapport est compris entre 0 et 1.

Si le rapport est de 0, alors aucun terme du label$_i$ n'est présent dans les labels des autres sujets pour ce groupement thématique. Par défaut, les réponses vides ou seulement identiques à la requête recoivent un score de 0 (nous les comptabilisons dans la moyenne des scores par catégorie). Lorsque le rapport est de 1, c'est que tous les termes du label ont été retrouvés dans les labels produits par les autres sujets pour le même groupement thématique. Le recouvrement est alors optimal. Cela permet d'avoir une idée de l'homogénéité des réponses produites par les sujets pour chaque regroupement.

Dans ce but, nous avons fait la moyenne des scores des labels pour chaque regroument évalué, ce qui permet d'évaluer l'homogénéité des labels proposés par chacun des sujets pour un même regroupement thématique. On peut ainsi voir l'accord des sujets face à la tâche proposée.

Résultats Les résultats de cette évaluation sont regroupés dans le tableau 7.8. Sur les 36 catégories évaluées, on dénombre 24 catégories présentant un score de recouvrement supérieur à 0,50 (soit 67% des catégories évaluées). Les tirets présents dans les cases du tableau 7.8 signalent que la catégorie visée n'existe pas pour la requête donnée. Les « 0 » signifient que l'évaluation n'a pas été possible (pas de réponses exploitables) pour une catégorie évaluée comme par exemple la catégorie CULTURES pour la requête *laetitia* (tableau 7.9). Les sujets ont donné peu de labels (seulement 5 répondants sur 20) et on voit que les labels ne partagent rien [5].

Le tableau 7.10 contient un exemple d'une catégorie ayant un score élevé, signifiant que les labels proposés par les utilisateurs pour les documents de

5. Le label *laetitia* est identique à la requête, donc on lui attribue un score nul par défaut. Cela permet d'en tenir compte lorsque l'on regarde l'homogénéité de chaque regroupement thématique.

cette catégorie se recouvrent fortement. On voit que le taux de réponse est bon, puisque l'ensemble des sujets a proposé un label. Les sujets ont formulé des labels très proches comme on peut le voir dans le tableau, les mots que l'on retrouve le plus sont *adolescent, hacker* et *français*. Ce recoupement de vocabulaire est quantifié par les scores de chaque label. Plusieurs d'entre eux obtiennent des scores de 1 ce qui signifie que tous les mots du label sont présents dans les autres labels comme pour *adolescent hacker wikileaks.*

	INT	SOC	CLT	ÉCO	POL	SPORT
afghanistan	0,68	0,60	0,74			
wikileaks	0,53	0,82	0,67	0,62		
berlusconi	0,61		0,78	0,27		
tunisie	0,58	0,37	0,19	0,79	0,64	
laetitia		0,79	0	0	0,60	
grève	0,27	0,64	0	0,69	0,49	0,83
égypte	0,68	0,65		0,21	0,62	
médicaments	0	0,40	0,73	0,62		
météo	0,74	0,80	0			

TABLEAU 7.8 : Score moyen de recouvrement des labels par catégorie et par requête

Labels	Score
la vie continue malgré cette affaire	0
laetitia	0
Meurtre de Laetitia, çà me parait plus important que Maison.....	0
RAP	0
dysfonctionnement justice	0

TABLEAU 7.9 : Laetitia - Catégorie CULTURES - Score moyen 0

Labels	Score	Labels	Score
adolescent hacker wikileaks	1,00	Wikileaks et les déboires d'un ado français avec internet	0,80
l'adolescent français et piratage	1,00	La face non officielle d'Internet	0,33
arrestation affaire Wikileaks	1,00	Wikileaks fait des émules encore adolescents	0,67
wikileaks ado francais pirate	1,00	adolescent hacker	1,00
adolescent arrété comme hacker	1,00	piratage de site	0,50
adolescents arrete	1,00	comment a proceder Wikileaks	0,33
ado français piratage informatique	0,75	adolescent hacker enquête	0,67
WikiLeaks : un adolescent français arrêté	1,00	wikileaks ado hacker	1,00
wikileaks hacker français	1,00	arrestation d'un adolescent français	1,00
Wikileaks pirate français	1,00	actualité d'un fait relaté par wikileaks	0,25

TABLEAU 7.10 : Wikileaks -catégorie SOCIÉTÉ - score moyen 0,82

Cette évaluation permet donc d'obtenir une image du comportement des utilisateurs vis-à-vis de la tâche et de montrer que la catégorisation des documents est compréhensible pour les utilisateurs. Cependant, cette méthode ne permet pas d'évaluer si la pluri-catégorisation est un bon moyen pour faire émerger la diversité portée par une requête d'un point de vue global. Pour cela, il faut évaluer le recouvrement des labels entre regroupements thématiques pour une requête donnée.

De plus, cette évaluation manuelle a permis de mettre en évidence certains phénomènes propres aux variations langagières et à la créativité des sujets. Nous avons relevé de nombreux labels ne se recoupant pas du point de vue de la forme, mais qui manifestent de forts recoupements sémantiques. Par exemple, pour la requête *afghanistan*, les sujets ont proposé comme labels *rôle*, *impact* ou encore *conséquences*. De même, on retrouve des termes comme *journaliste* et *reporter* dans les labels de la catégorie CULTURES pour la requête *af-*

ghanistan. Pour autant ces mots ne sont pas considérés comme équivalents par l'évaluation que nous avons pratiquée. Cette évaluation ne permet donc pas de mesurer le recouvrement sémantique des labels, ce qui minimise les recouvrements entre labels.

7.1.4.2 Évaluation des labels produits pour chaque groupement thématique pour une requête donnée

Une vision globale de l'évaluation des regroupements thématiques ne peut être obtenue que par une comparaison fine des labels commentant chacun des regroupements thématiques pour une requête donnée. Nous proposons donc de rechercher les termes de chaque label d'un regroupement thématique dans les labels des autres regroupements thématiques d'une requête. Cela nous permet de calculer la proportion de recouvrement des labels d'un regroupement thématique par rapport aux autres regroupements rattachés à la même requête. En effet, si un terme est présent, et ce plusieurs fois dans les labels de chaque regroupement thématique d'une requête, on peut supposer que ces regroupements sont assez semblables, ce qui peut questionner la pertinence de ce regroupement.

Calcul du recouvrement inter-catégories Pour ce faire, nous avons regroupé l'ensemble des labels produits pour un regroupement thématique sous la forme d'un « sac de mots ». Cette segmentation a été réalisée manuellement, du fait des labels parfois mal segmentés ou écrits avec des variantes orthographiques non classiques. La segmentation réalisée tient simplement compte des espaces. Nous considèrons comme étant un *type* les différentes formes d'un mot, de manière à regrouper les différentes orthographes sans toutefois regrouper les variantes de genre et de nombre. Une table de fréquence des termes est ensuite réalisée. Un filtrage est ensuite effectué avant le comptage des termes. Nous filtrons en effet les mots-outils tels que les prépositions, les pronoms et les déterminants qui apparaissent dans les labels. Nous filtrons également les mots identiques à la requête.

Proportion de mots pour une catégorie donnée qui sont partagés avec les autres catégories :

$$\frac{Token_{recouv} \times Type_{recouv}}{Token_{total} \times Type_{total}} \times 100$$

$Token_{recouv}$: nombre de tokens présents dans la catégorie évaluée et également présents dans une ou plusieurs catégories concomitantes

$Type_{recouv}$: nombre de types présents dans la catégorie évaluée et également présents dans une ou plusieurs catégories concomitantes

$Token_{total}$: nombre de tokens au total dans la catégorie évaluée

$Type_{total}$: nombre de types au total dans la catégorie évaluée

Ce calcul permet de pondérer le fait qu'un mot soit présent dans les labels des différents regroupements thématiques, en tenant compte de sa fréquence. En effet, il est intéressant de savoir si les labels des regroupements partagent plusieurs mots en commun tout en évitant les effets créés par un mot très fréquent de manière générale. Un faible taux de recouvrement indique alors que le regroupement thématique est spécifique et indépendant des regroupements concomitants. Un fort taux de recoupement des labels avec ceux d'une autre regroupement montre au contraire une faible « identité » de la catégorie thématique, celle-ci étant fortement semblable aux autres regroupements rattachés à la requête évaluée.

Résultats Les résultats de cette évaluation sont regroupés dans le tableau 7.11. Les tirets présents dans les cases du tableau 7.11 signalent que la catégorie visée n'existe pas pour la requête donnée. Les « / » signifient que l'évaluation n'a pas été possible (pas de réponses exploitables) pour un regroupement évalué. Enfin les « 0% » indiquent que le recouvrement est nul. Sur les 36 regroupements thématiques évalués, on dénombre 5 regroupements qui ne sont pas évaluables (marquées par des « / »). On compte 15 regroupements sur 32 évaluables, qui ont un taux de recouvrement compris entre 0% et 10% (marquées en bleu dans le tableau). À cela, on ajoute 7 regroupements qui ont un taux de recouvrement entre 10% et 20% (en rouge dans le tableau). Enfin, il y a 9 regroupements qui ont un taux de recouvrement supérieur à 20%. Nous considérons qu'au delà de 20% de recouvrement le regroupement contient des labels trop proches des labels des autres regroupements thématiques de la requête évaluée. Par exemple, on voit que la requête *afghanistan* est rattachée à 3 regroupements thématiques, pourtant 2 d'entre eux ont des taux de recouvrement élevés qui sont de 33,4% et 31,99% (INTERNATIONAL et SOCIÉTÉ, voir tableau 7.11). On peut donc dire que les regroupements INTERNATIONAL et SOCIÉTÉ ont reçu des labels qui partagent le même vocabulaire. Le troisième regroupement thématique CULTURES, dont les labels sont visibles dans le tableau 7.6, a un taux plutôt bas (9,76%).

	INT	SOC	CLT	ÉCO	POL	SPORT
afghanistan	33,40%	31,99%	9,76%			
wikileaks	41,36%	0,73%	4,03%	16,48%		
berlusconi	14,28%		10,24%	11,50%		
tunisie	53,08%	37,42%	27,77%	4,46%	8,57%	
laetitia		17,14%	/	/	9,70%	
grève	9,52%	4,21%	/	0,34%	0%	2,28%
égypte	18,97%	21,45%		31,67%	0%	
médicaments	/	3,33%	26,08%	7,66%		
météo	4,86%	13,13%	/			

TABLEAU 7.11 : Recouvrement inter-labels pour chaque regroupement thématique

L'analyse de ces résultats nous amène à constater que les observations sur la catégorie INTERNATIONAL se confirment. Comme nous l'avons vu dans le chapitre 6, cette catégorie peut être excessivement englobante. Les sujets ont proposé des labels pour les documents classifiés dans cette catégorie qui sont semblables aux labels d'autres regroupements thématiques. En effet, si on reprend l'exemple de la requête *afghanistan*, on voit très bien que les catégories INTERNATIONAL et SOCIÉTÉ auraient pu être fusionnées. Ceci s'explique par la nature de ces deux catégories qui étiquètent des documents qui parlent souvent de la même chose, mais d'un point de vue géographique différent. Nous avions soulevé ce problème dans le chapitre 5, lors de la présentation des catégories thématiques. Le phénomène s'accentue d'autant plus pour les requêtes *wikileaks* et *tunisie*.

On constate également que les requêtes de noms de pays ont des profils similaires vis-à-vis des taux de recouvrement des labels des regroupements thématiques (*tunisie, égypte* et *afghanistan*). Elles ont toutes une à deux catégories qui ont été labellisées avec des termes particuliers, comme pour la requête *tunisie* (catégorie ÉCONOMIE et POLITIQUE), et le reste des regroupements thématiques rattachés à ces requêtes ont des taux de recouvrement élevés. Par exemple, en ÉCONOMIE, les sujets ont donné des labels comme *tourisme perturbé tunisie* ou *rapatriment des touristes de la Tunisie* ; en POLITIQUE les labels sont différents : *nouveau gouvernement tunisie* ou *gouvernement tunisien.* Mais les autres regroupements qui étaient dominées par des documents sur le thème de la révolution tunisienne, ont été beaucoup plus difficiles à labelliser. Ceci était déjà apparu avec les résultats obtenus avec la première évaluation des labels, où les catégories INTERNATIONAL (0,58) SOCIÉTÉ (0,37) et

CULTURES (0,19) présentent des taux de recouvrement intra-catégorie moyens à faibles.

Les catégories rattachées à des requêtes NC ont des taux de recouvrement assez faibles dans l'ensemble (*grève, météo et médicaments*). Les sujets ont bien identifié les variations référentielles de ces requêtes. Pour la requête *grève*, ces résultats sont à pondérer grâce à la première évaluation. En effet, les catégories INTERNATIONAL et CULTURES ont été très peu labellisées.

Enfin, nous remarquons que les catégories ÉCONOMIE, POLITIQUE et SPORT ont des taux de recouvrement plus faibles. Nous avons vu dans la chapitre 6 que ces catégories étaient moins fréquentes et par conséquent moins généralistes. On peut supposer que les résultats seraient meilleurs avec une catégorisation plus fine qui compterait plus de thématiques.

7.1.5 Conclusion

Nous avons pu proposer deux mesures créées pour les besoins de l'évaluation. La plupart des regroupements thématiques a été évaluée positivement. Cependant, ces deux mesures montrent que si lorsqu'on considère les regroupements thématiques indépendamment des requêtes, on observe qu'ils sont majoritairement évalués positivement par les sujets. En revanche, si on considère les regroupements thématiques par requête, on voit que globalement chaque requête présente un regroupement thématique plus faible. En l'occurrence, ce regroupement évalué négativement est dans la majorité des cas une catégorie INTERNATIONAL. Par ailleurs, nous avons constaté que les sujets ont eu des difficultés à bien cerner les différents emplois de ces noms de pays lorsqu'ils apparaissaient dans des documents étiquetés dans des catégories trop généralistes comme INTERNATIONAL. Ces catégories rassemblent des informations variées et parfois identiques à d'autres catégories comme SOCIÉTÉ ou même CULTURES.

Nous avons affiné notre connaissance de la distribution des catégories thématiques. Elles sont pertinentes pour l'utilisateur lorsqu'elles permettent de révéler un évènement saillant. Elles permettent également de révéler des informations que les utilisateurs n'avaient pas vu comme certains ont pu nous le signaler lors de l'expérience.

7.2 Expérience 2 : l'utilisateur face à une tâche de catégorisation

Cette expérience se veut exploratoire. L'expérience 2 s'appuie en effet sur les techniques déployées pour l'expérience précédente. Elle permet aussi de la compléter. En effet, en demandant aux utilisateurs de pratiquer eux-mêmes des regroupements parmi les résultats d'une requête, on peut supposer qu'ils vont identifier la diversité portée par une requête.

Ainsi l'idée mise en pratique par cette expérience est de pouvoir évaluer la catégorisation des documents en elle-même et de confronter la catégorisation thématique « experte » à une catégorisation pratiquée par des non-experts.

Du point de vue de la mise en place de l'expérimentation, nous avons réutilisé les données évaluées en 7.1, dans le but de demander aux sujets de les regrouper librement. Cela permet d'avoir une approche ne nécessitant pas de nouveaux développements techniques et une adaptation minimale.

Nous allons donc tout d'abord décrire rapidement la mise en place de cette expérimentation. Nous exposons ensuite son déroulement puis les résultats de l'expérience 2. Enfin, nous terminons avec l'évaluation de ces résultats d'un point de vue quantitatif et qualitatif.

7.2.1 Mise en place de l'expérimentation

La mise en place de l'expérimentation reprend beaucoup d'éléments présentés en 7.1. Nous allons donc pointer les particularités de cette expérience, lorsqu'elles existent.

7.2.1.1 Choix des données

Les données utilisées et choisies pour cette expérience sont identiques à celles présentées en 7.1.1.1. Pour rappel, les requêtes évaluées sont présentées dans le tableau 7.1.

7.2.1.2 Les utilisateurs

Tout comme pour l'expérience de labellisation, les compétences attendues ont été consignées dans une fiche de renseignement que chaque sujet remplissait avant l'expérimentation. Les exigences minimales pour être sujet de l'expérimentation sont d'être un utilisateur averti des nouvelles technologies et de

s'intéresser régulièrement l'actualité. Les sujets choisis ne doivent pas avoir été influencés par l'expérience de labellisation, et ne doivent pas connaître les données. Le nombre choisi de sujets pour cette expérience est de 5. Un sujet va réaliser un pré-test en amont de l'expérience.

Pour rappel, le questionnaire de renseignement est en Annexe B.1 (page 210).

7.2.1.3 Traitements informatiques

Les traitements informatiques à mettre en place pour cette expérimentation sont beaucoup plus simples. En effet, il a suffi d'adapter l'interface conçue en php pour la précédente expérience de labellisation. De cette manière, il est possible de réutiliser les données regroupées par la catégorisation thématique, mais en laissant le soin au sujet d'opérer ces regroupements. Il est également possible de stocker les informations produites par les sujets de l'expérience, ce qui permet ensuite de les comparer aux classements automatiques de la catégorisation thématique.

L'interface dédiée au regroupement est ajoutée en post-traitement aux mêmes documents retournés par le moteur Terrier. Elle est visible dans la figure 7.12. Le nombre de documents affichés par catégorie est limité à 4, comme pour l'expérience de labellisation. Les paramètres sont identiques, à la différence que les documents sont affichés sans catégorie et dans un ordre aléatoire.

7.2.2 Déroulement de l'expérience 2

L'expérience 2 s'est déroulée du 20 au 24 février 2012 à Toulouse en suivant un protocole qui diffère en quelques points de celui de l'expérience de labellisation. Nous allons donc décrire le protocole en mettant en relief ces différences, nous poursuivrons ensuite en décrivant les sujets qui ont effectué l'expérience.

7.2.2.1 Protocole de l'expérience

Le scénario débute de manière similaire à la précédente expérience. Mais contrairement à l'expérience 1, nous interagissons avec le sujet tout au long de la passation. Le sujet ne se voit imposer aucune limite de temps. Voici une vue synthétique du scénario :

1. Remplissage du questionnaire par le sujet
2. Présentation de l'expérience au sujet
3. Consigne

4. Passation de l'expérience de manière interactive
5. Expérience terminée - Entretien de « débriefing »

Les étapes 1 et 2 sont semblables à celles proposées dans l'expérience de labellisation présentée dans la section 7.1.

Lors de l'étape 3, la consigne est expliquée à l'utilisateur. Il a devant lui la requête *mam* et les résultats retournés à cette requête sous format de liste (voir figure 7.12 page suivante). Nous lui demandons de regarder attentivement les titres de documents présentés en résultat à la requête proposée dans la partie haute de l'interface et nous lui indiquons qu'il peut avoir accès au texte du document en passant la souris dessus. Voici la consigne proposée au sujet :

> *Votre tâche va consister à regrouper, réorganiser ces résultats, c'est-à-dire que vous devez rassembler les documents entre eux selon un critère qui semblera pertinent.*
>
> *Pour cela, un champ libre est placé à côté de chaque document. Vous indiquerez par un signe distinctif commun (un chiffre) que vous souhaitez regrouper tel document avec tel autre document présent dans les résultats.*
>
> *Enfin, vous pouvez faire un minimum de deux « paquets » et au maximum cinq « paquets ».*

Il est précisé à l'utilisateur que les chiffres ne seront pas considérés de façon hiérarchique, ils servent juste à distinguer les « paquets » entre eux.

Pour l'étape 4, l'expérience se déroule de manière interactive, c'est-à-dire que nous demandons aux sujets d'expliciter leurs choix en matière de regroupement. Nous opèrons une prise de note active lors de chaque expérimentation. Une fois l'expérience finie sur les 9 requêtes du panel, nous proposons un moment d'échange libre avec le sujet pour recueillir ses impressions par rapport à l'expérience (étape 5).

Questions :
– *Est-ce que vous avez trouvé ça complexe ?*
– *Est-ce que vous avez utilisé vos connaissances sur l'actualité ?*
– *Quelle a été votre stratégie ?*

Les questions sont plus simples et plus courtes que dans l'expérience 1. Cela est dû au fait que l'expérience se déroule de manière interactive et, par conséquent, nous avons l'occasion de poser des questions au fur et à mesure. Mais elles servent toujours à clôturer la séance d'expérimentation avec le sujet et à lui permettre de poser à son tour des questions.

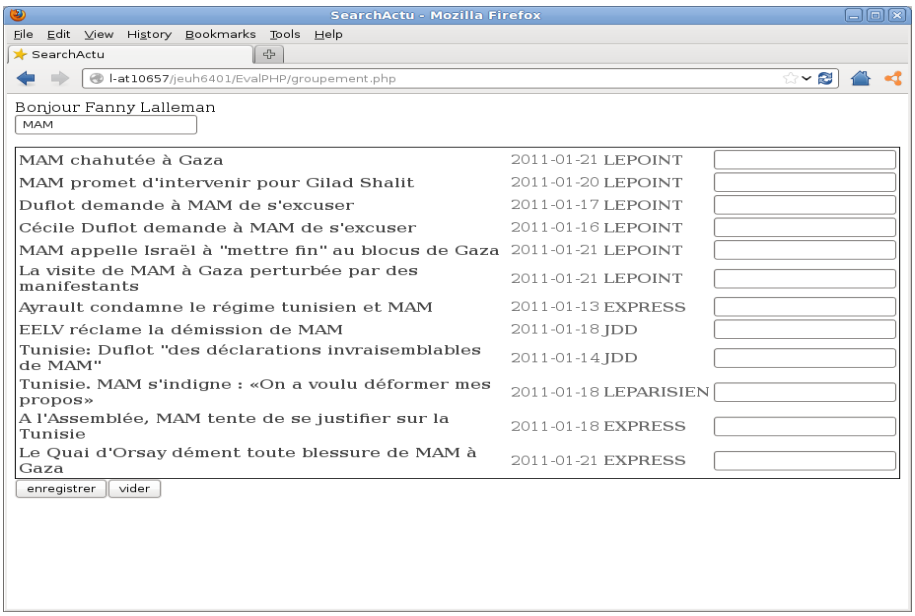


FIGURE 7.12 : Interface de l'expérience 2 de catégorisation

7.2.2.2 Les sujets de l'expérience

Cette expérience étant beaucoup plus exploratoire mais également beaucoup plus longue à réaliser, elle a été effectuée par un nombre plus modeste de sujets. Un sujet a réalisé un pré-test et l'expérience en elle-même a été réalisée par 5 sujets (3 hommes et 2 femmes). Les sujets ont été recrutés sur la base du volontariat parmi mes connaissances. Quatre sujets ont déclaré s'informer quotidiennement et un sujet a déclaré s'informer au moins 2 à 3 fois par semaine. Les usages en matière d'information different des usages des sujets de l'expérience de labellisation comme on peut le voir sur la figure 7.13. Les journaux télévisées sont la première source d'information des sujets, 100% des sujets ont déclaré s'informer via les journaux télévisés. Les autres sources d'information sont à égalité : radio, presse numérique et plates-formes d'actualité.

Mais malgré ces observations, on retrouve une tendance à la multiplication des sources d'information comme pour les sujets de l'expérience de labellisation. En effet, en moyenne un sujet dit consulter 3,4 sources différentes pour s'informer.

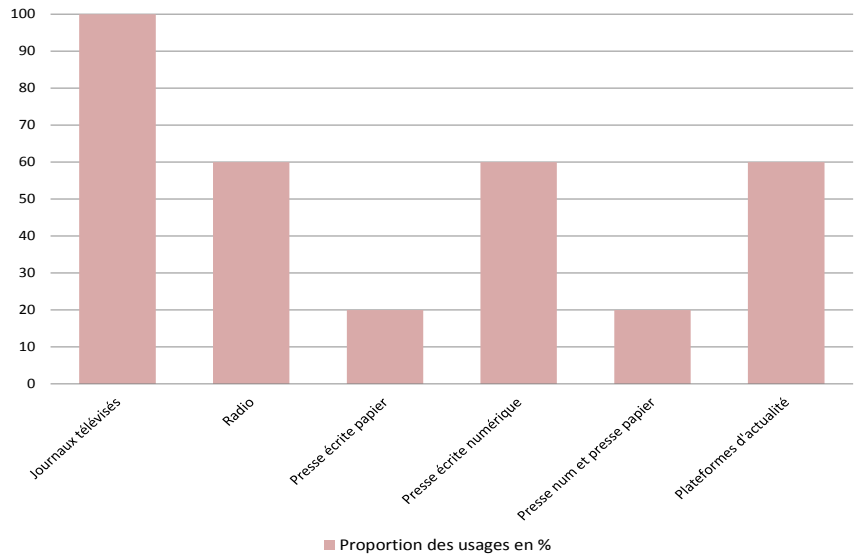


FIGURE 7.13 : Usages en matière d'accès à l'information des sujets de l'expérience de regroupement

7.2.3 Résultats

Les sujets ont effectué l'expérience en moyenne en 53,6 minutes. Toutefois, un sujet a mis beaucoup plus de temps que les autres sujets pour accomplir les regroupements demandés (116 minutes). Les quatre autres sujets ont mis en moyenne 38 minutes pour effectuer cette tâche.

Les sujets ont expliqué avoir changé de stratégies dans leur tâche de regroupement. En effet, par exemple le sujet user02 a déclaré avoir organisé au départ les documents en fonction de leurs dates, puis ensuite en fonction de leurs contenus. Le nombre de regroupements semble aussi influer sur le choix de leur stratégie : plus le nombre de groupes possible augmente, moins le regroupement « thématique » est possible. Le regroupement s'oriente vers un rapprochement de type d'article ou de sujets comme *faits-divers médicaux* ou *articles de société*.

Cette tâche a été évaluée par les sujets comme étant difficile. Certains sujets ont déclaré qu'ils auraient fait autrement s'ils devaient refaire les regroupements. Les sujets ont déclaré avoir fait appel à leurs connaissances quand ils se souvenaient des informations.

Le nombre de regroupements opérés par les sujets oscille entre 3 et 5 selon les requêtes considérées comme on peut le voir dans le tableau 7.14. Le nombre de regroupements proposés par les sujets est comparable à celui produit automatiquement via les catégories thématiques, à l'exception de la requête *grève* qui totalise 6 catégories alors que les sujets ne pouvaient en faire que 5 au maximum à cause d'une erreur.

Requêtes	Sujets	Catégories
afghanistan	3,4	3
wikileaks	4	4
berlusconi	3	3
tunisie	4,8	5
laetitia	4,2	4
grève	5	6
égypte	3,8	4
médicaments	4,2	4
météo	4,4	3

TABLEAU 7.14 : Nombre moyen de regroupements par requêtes

7.2.4 Évaluation

Afin d'évaluer la classification établie par les sujets de l'expérience et de la confronter à la catégorisation automatique, nous proposons d'utiliser le Rand Index (RI). C'est une méthode statistique établie par Rand (1971) pour comparer la similarité de deux méthodes de classification. Il existe également une version ajustée du Rand Index (ARI) qui permet de corriger la mesure obtenue par rapport à un résultat obtenu grâce au hasard.

La méthode du Rand Index considère les documents 2 à 2 et permet de dire s'ils sont traités d'une manière identique par les deux types de catégorisation (humaine et automatique) ou non. On obtient alors un pourcentage de paires traitées de manière identique. Le Rand Index est une mesure de précision dont le maximum est 1 (pour un classement identique).

Nous avons rassemblé les résultats de cette évaluation par ARI et RI dans le tableau B.11 (Annexe B.2). Ce tableau détaille les résultats sujet par sujet et propose une moyenne des pourcentages de précision obtenus pour chacune des requêtes évaluées.

Ces résultats montrent que globalement les sujets ont réussi à établir des classements intéressants, qui ne sont pas si éloignés de la classification automa-

tique. Les valeurs de ARI nous permettent aussi de diagnostiquer que les documents rattachés à trois requêtes ont posé les mêmes problèmes à l'ensemble des sujets. Ce sont les requêtes *berlusconi*, *médicament* et *egypte*. Ces résultats s'expliquent par le fait que ces requêtes ont ramené une trop grande diversité de documents. Par exemple, la requête *berlusconi* ramène des documents très éloignés en terme de contenu, les documents concernant ses activités de président du conseil et des faits-divers où il est impliqué. Les résultats sont visibles dans le tableau 7.16. Or dans cet exemple, les sujets se demandaient s'il fallait plutôt privilégier les évènements ou les thématiques abordées par les articles. Cela se matérialise par la constitution d'un groupe intitulé *croissance italienne* à côté d'un groupe appelé *affaire ruby*.

Nous remarquons toutefois une requête qui a été classifiée de manière quasiment semblable à la catégorisation thématique par l'ensemble des sujets. C'est la requête *météo*. Elle obtient en moyenne un RI de 0,85 et un ARI de 0,74 (tableau 7.15). Cette requête, dans la période temporelle évaluée, ramenait des documents qui avaient plusieurs sujets : un film sur une miss météo, fan de Marilyn (Poupidou) et divers documents sur la météo à la fois à l'international et en France. Le tableau 7.17 reprend les réponses des sujets. On trouve dans ce tableau le nombre de groupes constitués par le sujet et les différents labels qu'ils ont donnés à chacun des groupes. Nous constatons que 4 des sujets ont décidé de constituer un groupe à propos du film Poupidou. Le reste des groupes constitués ont en commun le même type de documents (météo, évènements climatiques ou naturels). Ce type de classement se rapproche alors du classement automatique : INTERNATIONAL, SOCIÉTÉ et CULTURES.

sujet	ARI	RI
user2	0,86	0,86
user4	0,53	0,86
user5	0,86	0,86
user6	0,71	0,82
user7	0,78	0,85
moyenne	0,74	0,85

TABLEAU 7.15 : Résultats ARI et RI pour la requête *météo*

Les autres requêtes évaluées présentent des valeurs de précision correctes comme par exemple la requête *afghanistan* (tableau 7.18), où l'on voit trois thématiques se détacher pour l'ensemble des sujets : al-quaïda, les otages français et les soldats. Ces thématiques correspondent à des évènements bien identifiés que la requête a ramenés en résultats. Nous voyons que les évènements

sont plus facilement identifiables à partir de la classification humaine, en comparaison à la classification thématique qui est moins fine.

Sujets	Groupe	Labels des groupes
user2	3	[l'économie] [justice contre berlusconi] [les réactions de berlusconi]
user4	2	[affaire ruby] [croissance italienne]
user5	3	[justice, parquet] [mesures éco] [politique&berlusconi]
user6	3	[berlusconi gouvernement] [ruby] [menaces démission berlusconi]
user7	4	[rubygate] [action de berlusconi] [manif] [berlusconi ne veut pas démissionner]

TABLEAU 7.16 : Exemple des classements réalisés par les sujets (requête *berlusconi*)

Sujets	Groupe	Labels des groupes
user2	4	[poupidou] [météo hors de France] [météo en France] [éclipse]
user4	5	[poupidou] [les DOM-TOM] [l'éclipse+Rio] [intervention florence] [météo France]
user5	4	[poupidou] [climat hors métropole] [climat en métropole] [éclipse]
user6	5	[poupidou] [tempêtes tropicales] [au secours la France] [Rio] [éclipse]
user7	4	[France métropolitaine] [outre-mers] [actu secondaire] [l'étranger]

TABLEAU 7.17 : Exemple des classements réalisés par les sujets (requête *météo*)

Sujets	Groupe	Labels des groupes
user2	4	[soldats tués] [les otages décompte] [obama, sujet international] [?]
user4	4	[al quaïda] [otages journalistes] [soldats morts] [ACMI dernières news]
user5	2	[otages] [conflit avec militaires]
user6	3	[otages] [soldats] [al-quaïda]
user7	4	[soldats français] [otages journalistes] [otages en général] [al-quaïda]

TABLEAU 7.18 : Exemple des classements réalisés par les sujets (requête *afghanistan*)

7.2.5 Conclusion de l'expérience 2

Cette deuxième expérience qui se voulait exploratoire et qualitative nous apprend plusieurs choses. Les sujets ont été confrontés à une tâche demandant de la concentration sur une durée assez prolongée, des compétences (compréhension et classement de documents) et une bonne connaissance de l'actualité.

L'évaluation de la tâche de regroupement montre que la classification automatique se rapproche de la classification manuelle pour la plupart de nos requêtes. Seuls les résultats de deux requêtes se sont montrés particulièrement difficiles à classifier comme nous venons de le voir. En effet, les sujets ont été mis en difficulté par la qualité des documents ramenés en résultats.

7.3 Conclusion

La réalisation des deux expériences a mis en lumière certains éléments du point du vue des utilisateurs et de leurs usages. Particulièrement à propos des catégories thématiques, qui ont été jugées par les sujets de l'expérience 1 comme étant « correctes ». Les catégories thématiques remplissent donc leur rôle dans l'expérience 1. Plusieurs d'entre eux ont également souligné les ressemblances avec les sites d'actualité.

Les deux expériences ont montré leur complémentarité : la première a permis de tester la compréhension du dispositif et sa validité, alors que la seconde a montré que le classement produit par des personnes pouvaient s'approcher de la classification automatique. Cette double démarche nous permet d'identifier certains problèmes et certaines requêtes dont les résultats ramenés ne sont

pas classifiés de manière satisfaisante. Ces requêtes ont en particulier soulevé le problème de la granularité des informations ramenées par le moteur de recherche. Ce problème de granularité serait peut être minimisé en apportant une amélioration à l'interface de recherche, en ne figeant pas le nombre de documents retournés par catégorie. À la place, nous proposons d'appliquer un filtre par degré de similarité avec la requête.

Ces expériences nous permettent de confirmer l'intérêt des catégories thématiques. Les utilisateurs ont pu comprendre et formuler la diversité qui leur était proposée. Nous avons également pu observer le comportement des utilisateurs vis-à-vis de requêtes entraînant une diversité dans les résultats de recherche. Cela a fait apparaître que les utilisateurs peuvent avoir une perception de l'ambiguïté qui diffère de celle qui est renvoyée par la base documentaire.

Par conséquent, il nous paraît intéressant d'explorer d'autres éléments contextuels à notre disposition. Nous pouvons en effet nous demander en quelle mesure les indices peuvent être complémentaires dans notre contexte de RI. Pour cela, nous allons effectuer un examen qualitatif et exploiratoire de ces indices dans le chapitre suivant.

Chapitre 8

Examen qualitatif d'indices contextuels complémentaires

L'utilisation des indices contextuels spécifiques à l'application tels que la catégorisation thématique nous a permis de comprendre certaines manifestations de l'ambiguïté. Ce chapitre est l'occasion d'ouvrir des perspectives sur le traitement de l'ambiguïté au niveau de notre application mais aussi au niveau de la RI en général.

Les chapitres 6 et 7 nous ont permis de voir à la fois les points forts mais aussi les points faibles des indices contextuels tel que la catégorisation thématique. Nous pensons que la combinaison d'indices contextuels supplémentaires offre la possibilité de disposer d'un faisceau d'informations. Nous faisons l'hypothèse que ce faisceau d'information contextuelle va nous permettre de considérer le comportement d'une requête d'une manière globale. Dans ce but, nous considérons des indices contextuels supplémentaires provenant des utilisateurs. Comme nous l'évoquions dans le chapitre 3, les requêtes sont la principale trace du contexte utilisateur. Deux indices sont particulièrement intéressants et réutilisables : les requêtes répétées et les requêtes reformulées. Nous complétons cette exploration des indices contextuels par l'apport des co-occurrences. Cet indice est un appui supplémentaire pour analyser le contexte linguistique présent de la base documentaire.

Dans ce chapitre, nous commençons par explorer le potentiel de ces deux derniers indices non exploités jusqu'à présent en les adaptant aux données dont nous disposons. Nous proposons ensuite de nous appuyer sur l'ensemble de ces indices à notre disposition pour pouvoir développer une étude de cas, dans laquelle nous analysons plusieurs cas de requêtes qui nécessitent d'avoir des modes de présentation de résultats diversifiés. En effet, les causes de

l'ambiguïté ou de l'indétermination qui touchent ces requêtes ne sont pas les mêmes, et par conséquent c'est cette diversité que nous allons illustrer par cette analyse.

8.1 Un indice contextuel : les versions étendues des requêtes

Nous avons vu dans les chapitres 2 et 3 (cf. 3.2.3.2) que la reformulation des requêtes peut avoir un intérêt pour désambiguïser certains mots des requêtes (Shen et Zhai, 2003). En effet, le repérage des reformulations d'une requête permet d'étendre le contexte de la requête, et de préciser grâce aux requêtes reformulées le besoin d'information, voire de révéler la diversité d'une requête (Santos *et al.*, 2010a).

Bien que les informations concernant l'utilisateur soient très incomplètes dans les cas de nos données, les requêtes sont en elles-mêmes des traces de la diversité des modes d'expression qui sont utilisés au fil des recherches (Song *et al.*, 2009; Jansen *et al.*, 2007). La façon dont une requête a été reformulée, ou en tous cas (en l'absence d'information sur l'identité de l'utilisateur), les différentes manières dont un terme a été utilisé dans une série de requêtes, fournissent des indices sur la diversité des points de vue exprimés par les utilisateurs.

La version « étendue » d'une requête courte peut nous informer sur les types de spécification possible d'une requête sémantiquement large. Toutefois, deux requêtes en apparence liées du point de vue de leur forme peuvent n'avoir aucun autre lien comme *transports* (mai) qui est étendue en *transports 40 tonne a 44tonnes* ou encore *agriculture* (mai) étendue en *http agriculture gouv fr beneficiaires pac* ou *sarkozy salon de l agriculture.*

8.1.1 Mesurer la capacité d'extension d'une requête courte

Les requêtes étendues ont été recensées sur 4 sous-corpus du corpus *2424reqFréquentes* : mai, juin, octobre et décembre 2010. Nous avons recensé les extensions de l'ensemble de ces requêtes, soit environ 195 requêtes. Les extensions ont été recherchées sur l'ensemble des logs des mois correspondants. Nous avons également inclus dans cette étude les requêtes multi-termes, qui avaient été exclues de l'étude de l'ambiguïté. C'est en effet également l'occasion de voir si ces requêtes sont étendues.

Nous avons extrait des logs de requêtes 2424actu les requêtes contenant la *requête initiale+x* (exemple 1) ou *x+requête initiale* (exemple 2), ainsi que les fréquences d'apparition dans le log de chaque extension.

(1) requête initiale : *gouvernement*, requête initiale + x = *gouvernement belgique*

(2) requête initiale : *facebook*, x + requête initiale = *apéro facebook*

60% des requêtes de ce corpus ont au moins une version étendue, et 30% n'en ont aucune. Parmi les requêtes qui n'ont aucune extension, on retrouve des requêtes multi-termes comme *loi voile, sport tennis* ou *proxenetisme equipe de france* mais également des requêtes NPP comme *johnny hallyday* ou *roman polanski* (en mai 2010).

Enfin, un certain nombre d'extensions ne sont que des modifications d'une lettre ou l'ajout d'un déterminant comme *audrey pulvar*, dont la version étendue est *audrey pulvard*. Ces modifications sont peu intéressantes et sont souvent le signe de fautes d'orthographe.

Nous proposons à présent d'examiner les requêtes étendues en fonction de la productivité puis ensuite d'examiner les 30% de requêtes qui n'ont pas de versions étendues dans nos corpus. Nous voulons voir pourquoi certaines requêtes ont des formes étendues et d'autres non.

8.1.1.1 Les requêtes étendues

La mesure du degré d'extension d'une requête nous permet d'évaluer le volume des versions étendues vis-à-vis de la requête initiale. Pour cela, nous avons simplement cherché à calculer un ratio qui compare la fréquence de la requête initiale + ses versions étendues, à la fréquence des versions étendues de la requête. On peut formaliser ce ratio par la formule suivante :

$$Ratio = \left(\frac{f_{rE}}{f_{rI}+f_{rE}}\right)$$

où f_{rE} : fréquence des requêtes étendues et f_{rI} : fréquence de la requête initiale

Ce ratio s'étend entre 0 et 1. Nous allons tout d'abord examiner les requêtes ayant un ratio important entre 1 et 0,5 puis un ratio plus modéré entre 0,5 et 0, et enfin les requêtes n'ayant pas de versions étendues.

Les requêtes fortement étendues Nous observons grâce à ce ratio, le cas intéressant des requêtes dont la version initiale a une fréquence inférieure à ses versions étendues. Ces requêtes représentent à peine 4% du corpus de requêtes étudiées. Par exemple, la requête *retraite* dans le tableau 8.1 est beaucoup moins fréquente que ses versions étendues. Cette requête apparaît dans

de nombreuses requêtes plus longues : *reforme retraites* (fréquence de 3 840), *reforme retraite* (fréquence de 110), *reforme des retraites* (fréquence de 74), *regime de retraite* (fréquence de 7), *depart en retraite* (fréquence de 7). Elle est également étendue avec des changements mineurs comme *retraite 62* ou *retraites*. Tout en étant fréquente, la requête *retraite* est supplantée en popularité par la requête *reforme retraites*, ce qui explique un ratio élevé.

Les requêtes regroupées dans le tableau 8.1 ont un ratio supérieur à 0,5 pour des raisons comparables à la requête *retraite*, c'est-à-dire qu'elles sont présentes dans une requête étendue qui les supplante en popularité. Ainsi la requête *gouvernement* a pour version étendue *gouvernement belgique* (avec une fréquence de 381), ou encore la requête *facebook* qui se reformule en *apero facebook* au mois de juin (fréquence de 4415), associée au mot *apéro*, elle désigne alors une série d'évènements.

Corpus	requête rI	exemple rE la plus fréquente	freq rI	freq rE	freq rI+rE	ratio
mai	*facebook*	*apero facebook*	926	2655	3581	0,74
juin	*retraite*	*reforme retraites*	176	4140	4316	0,95
juin	*facebook*	*apero facebook*	251	4416	4667	0,94
juin	*xynthia*	*tempete xynthia*	128	176	304	0,57
juin	*gouvernement*	*gouvernement belgique*	307	391	698	0,56
oct	*reforme retraite*	*reforme retraites*	154	3492	3646	0,95
oct	*mineurs chili*	*mineurs chili fr*	533	5560	6093	0,91
oct	*otages*	*otages niger*	159	575	734	0,78

TABLEAU 8.1 : Requêtes (rI) dont les versions étendues (rE) sont plus fréquentes (ratio > 0,5)

Pour ces requêtes, on constate donc que le ratio élevé est surtout associé à la présence d'une forme qui concentre la majorité des extensions.

Les requêtes peu ou moyennement étendues Les requêtes ayant un ratio entre 0,5 et 0,01 représentent 27% des requêtes étudiées et 38% ont un ratio compris entre 0,01 et 0. Les résultats sont détaillés dans le tableau 8.2 qui regroupent le nombre de requêtes étendues présentant un ratio compris dans la fourchette considérée. Ce sont les corpus des mois d'octobre et décembre qui contiennent le plus de requêtes étendues. Nous détaillons quelques exemples illustrant ce cas de figure.

Corpus	Total
ratio 0,5 et 0,01	54
ratio > 0,01	75

TABLEAU 8.2 : Nombre de requêtes avec un ratio > 0,5

La requête *iphone* (corpus de juin) présente un ratio de 0,46. Ce ratio élevé s'explique par une fréquence de base peu élevée associée à un nombre relativement important d'extensions en proportion. Les versions étendues de cette requête (tableau 8.3) sont principalement une spécification du modèle d'iphone en l'occurrence l'iphone 4. Les autres requêtes étendues sont des hapax.

Fréquence	Requêtes étendues
65	*iphone 4*
17	*iphone4*
10	*iphone4*
2	*iphone 4 g*
1	*iphone 4g date de sortie - iphone 4 sfr - sortie iphone 4 - iphone 4 antennes - iphone capte pas - iphone os4 - iphone probleme - nouveau iphone - prix iphone 4*

TABLEAU 8.3 : Les extensions de la requête *iphone* (juin)

La requête *carla bruni* en juin a un ratio inférieur à celui de la requête *iphone* puisqu'il est de 0,13. Mais comme la requête précédente, une seule forme de requête étendue est fréquente : *carla bruni photos.* Comme on peut le voir dans le tableau 8.4, seulement une autre version étendue existe dans ce corpus, *carla bruni a doha.* Cette extension faisant référence à un voyage présidentiel où Carla Bruni a accompagné son mari Nicolas Sarkozy est peu fréquente.

Dans ces deux cas de requêtes étendues, les extensions se concentrent sur une seule forme, l'éparpillement est faible.

Fréquence	Requêtes étendues
65	*carla bruni photos*
3	*carla bruni a doha*

TABLEAU 8.4 : Les extensions de la requête *carla bruni* (juin)

La requête *corée* (décembre 2010) est un exemple intéressant parce que malgré un ratio très bas de 0,025, elle présente des versions étendues qui méritent attention. Comme on peut le voir dans le tableau 8.5, les extensions les plus fréquentes sont *coree du nord*, suivi par *coree du sud*, renvoyant aux emplois homonymiques de la requête *corée* décrits dans le chapitre 6. Nous voyons également une série d'extension hapax à propos du conflit opposant la Corée du nord à la Corée du sud (*nouvelles menaces coree* ou *riposte coree du sud*), ce qui n'est pas le cas pour les hapax *maillot de bain coree*, *musique coree*, *pop coreene* et *coree education*.

Fréquence	Rêquetes étendues
40	*coree du nord*
5	*coree du sud*
2	*2 chine coree*
1	*coree d - coree de nord - coree nord - crise coree du nord - defense de la coree - nouvelles menaces coree - riposte coree du sud - iran coree - japon coree - guerre de coree - maillot de bain coree - musique coree - pop coreene - coree education*

TABLEAU 8.5 : Les extensions de la requête *corée* (décembre)

Parmi les requêtes ayant un très faible ratio, on observe certaines requêtes qui ont été étiquetées comme ambiguës par Wikipédia dans le chapitre 6, comme la requête *société* (ici au mois de mai). Elle a été également pluri-catégorisée. Or elle a une seule forme étendue dans notre corpus : *société générale*. Cette quasi absence de requête étendue s'explique peut-être par le fait que cette requête aurait pu avoir une fonction navigationnelle et non pas informationnelle. L'interface de 2424actu contient en effet une catégorie du même nom (cf. chapitre 4, figure 4.1 page 70). La requête *international* présente le même type de situation et n'a pas non plus de formes étendues dans notre corpus.

On retrouve également des requêtes comme *alain juppe* (ratio de 0,005 en décembre), *pakistan* (ratio de 0,004 en octobre) ou encore *prince william* (ratio de 0,001 en décembre). Ces requêtes ont une ou deux requêtes étendues au maximum comme on peut le voir dans le tableau 8.6 pour la requête *prince william*.

Fréquence	Requêtes étendues
1	*nouvelles photos du prince william et de kate*
1	*mariage prince william*

TABLEAU 8.6 : Les requêtes étendues de la requête *prince william* en (décembre)

8.1.1.2 Les requêtes sans extension

30% des requêtes étudiées n'ont aucune extension, ce qui représente une proportion de requêtes importante. Les résultats détaillés sont exposés dans le tableau 8.7.

Corpus	mai	juin	oct	déc
Sans extension	12	17	12	17

TABLEAU 8.7 : Nombre de requêtes n'ayant pas de formes étendues

Les requêtes sans extension comptent toutes plus de 2 mots, à l'exception de *berlusconi* et *golf* (en octobre). Le tableau 8.8 rassemble des exemples de requêtes sans extension. La plupart de ces requêtes n'a pas été catégorisée par notre procédure parce qu'elles contiennent plusieurs termes comme *proxenetisme equipe de france.*

Corpus	Requêtes étendues
juin	*proces kerviel - maree noire etats unis - sport tennis*
mai	*proxenetisme equipe de france - crise grece - seisme chine - identite nationale*
oct	*mineurs chili fr - otages niger - laurent fignon*
déc	*arnaud montebourg - tony parker - nicolas dupont aignan*

TABLEAU 8.8 : Exemples de requêtes sans extension

8.1.2 Conclusion

L'étude des capacités d'extension des requêtes de notre corpus révèle que c'est un indice complexe à interpréter. En effet, le taux d'extension d'une requête n'est pas forcément signe qu'une requête est ambiguë. Nous avons même vu que des requêtes pouvant avoir potentiellement des fonctions variées (navigationnelle ou informationnelle) ne présentaient quasiment aucune version étendue dans le corpus. Nous avons également observé que les extensions pouvaient aussi se concentrer sur une seule forme.

Par ailleurs, notre corpus d'étude était de taille modeste et les parcours de recherche n'étaient pas identifiables. Ce sont des élements qui limitent la portée de cet indice. La présence des parcours aurait pu permettre d'envisager la reformulation comme un indice d'ambiguïté autonome, l'étude des versions

étendues ne donnant pas le même type d'information. Mais pour cela il faudrait disposer des parcours de recherche complets des utilisateurs. Ce type d'information permettrait de tester des solutions de personnalisation pour résoudre les problèmes d'ambiguïté. On pourrait ainsi savoir ce qu'il se passe lorsqu'un utilisateur a soumis une requête potentiellement ambiguë et qu'il n'a pas fait de reformulation. Deux situations sont possibles : l'utilisateur poursuit sa recherche, ou l'utilisateur stoppe sa tâche de recherche, découragé par les résultats.

8.2 Un indice contextuel : les cooccurrences

Les cooccurrences sont le dernier indice utilisé. Ce procédé est couramment utilisé pour étudier la variation sémantique en diachronie (Picton, 2009) ou la polysémie (Yarowsky, 1995; Turney, 2004). Les cooccurrences sont utilisés pour effectuer de la désambiguïsation sémantique (Jacquet et Venant, 2005).

L'intérêt d'un tel indice réside principalement dans sa capacité à faire émerger des différences d'emploi d'un mot en contexte (Habert *et al.*, 2005). En effet, les cooccurrences permettent de répérer les mots « mouvants ». Notre but est justement d'identifier si un mot utilisé comme requête est mouvant dans les documents d'actualité. Les cooccurrences sont censées nous permettre de détecter une hétérogénéité sémantique (Habert *et al.*, 2005). Les classes de mots se constituent autour d'éléments saillants du corpus (Jacquet et Venant, 2005). Les analyses peuvent être réalisées à partir des rapports de dépendances syntaxiques entre un mot et son contexte ou à partir des rapports de voisinage distributionnel.

L'intérêt de cette analyse distributionnelle dans notre contexte d'étude est de pouvoir identifier des liens forts entre la requête et des cooccurrents qui permettent de dégager un ou des comportements sémantiques de cette requête. L'approche que nous avons menée est néanmoins restée exploratoire : elle est appuyée sur un examen manuel et n'a pas permis de déboucher sur un score d'ambiguïté potentielle.

Nous procédons ici à une analyse des cooccurrents de surface à l'aide de l'outil Antconc (Anthony, 2011) sur le corpus 2424 non lemmatisé, en utilisant la mesure d'information mutuelle (notée IM). Le contexte considéré est une fenêtre de 3 mots avant et après l'unité étudiée. Nous sélectionnons automatiquement les cooccurrents les plus représentatifs en fonction de leur score d'information mutuelle et ayant une fréquence supérieure à 3 dans le corpus. Nous signalons que l'IM attribue des scores forts à des formes atypiques comme les erreurs orthographiques. De plus, notre analyse demande une interprétation manuelle

car nous ne procédons pas à un dégroupement automatique dans cette étude. La démarche est donc limitée.

Nous allons présenter ici des exemples montrant ce que peut apporter l'analyse des cooccurrents à l'analyse des requêtes mais également les problèmes rencontrés. Les exemples choisis sont des requêtes de novembre 2010 : *ministre* et *international*. Les cooccurrences seront particulièrement utilisées dans la section suivante en 8.3, lors de la combinaison des indices pour l'analyse de cas.

Le tableau 8.9 montre les collocats les plus proches statistiquement du mot *ministre*. Cet exemple met en évidence des collocats qui sont similaires par leur type. Ainsi, nous trouvons des collocats qui sont principalement des noms de personnes : *Leterme, Singh, Papandreou, Nazif, Naoto* ou *Manmohan*. Ces collocats sont le signe que ce mot a besoin d'être spécifié en contexte, particulièrement pour des précisions référentielles. Nous signalons également un problème récurrent, un mot mal orthographié a également été repéré (*remier* pour « premier »). Ces phénomènes sont accentués par la nature du corpus du document. En effet, les articles de presse sont parfois des reprises d'une publication de l'AFP. Dès lors que l'AFP produit un document avec des erreurs, plusieurs médias peuvent reprendre la publication en l'état. Cela crée une redondance des informations mais également des erreurs.

Collocats de *ministre* (Nov 2010)
remier (IM :9,8), Leterme (IM :9,8), Vice (IM :9,5), Singh (IM :9,2),Ressources (IM :9,2), Papandreou (IM :9,2), Nazif (IM :9,2), Naoto (IM :9,2), Manmohan (IM :9,2), Initie (IM :9,2)

TABLEAU 8.9 : 10 collocats les plus proches statistiquement (IM)

Pour le mot *international* (tableau 8.10), on trouve des collocats de type différent mais qui occupent le même type de fonction. Ils viennent spécifier le mot en question et, dans plusieurs cas, ils vont former un terme comme avec *Amnesty, Fonds* ou *Transparency* : Amnesty International, Fonds Monétaire International (FMI), Transparency International (ONG anti-corruption). On voit également un acronyme parmi les collocats, *TI*, pour « tribunal international ». C'est donc un mot qui est faiblement autonome dans les documents et qui est prédisposé à une association lexicale pour former des entités nommées. Ce constat est en décalage avec les observations menées sur la requête *international* qui n'a aucune version étendue. Cela laisse supposer que ce mot non autonome dans les documents forme une requête qui n'a pas besoin d'être précisée.

Collocats de *international* (Nov 2010)
Monétaire / monétaire (IM :15,1), Consortium (IM :15,1), Amnesty (IM :14,9), Fonds (IM :14,7), Transparency (IM :14,6), Transparence (IM :14,3), consortium (IM :14,3), TI (IM :13,6), refinancer (IM :13,6), tollé (IM :13,3)

TABLEAU 8.10 : 10 collocats les plus proches statistiquement (IM)

Cette étude limitée fournit des résultats difficiles à interpréter en synchronie. Dans la section suivante, on verra qu'en diachronie on peut repérer plus facilement des constantes. En effet, les cooccurrences sont seulement utilisées comme un appui contextuel supplémentaire car la méthodologie d'extraction reste limitée. Cet appui nous permet d'observer des décalages entre le contexte présent dans la base documentaire et le contexte utilisateur.

8.3 Combinaison des indices contextuels : analyse de cas

Nous proposons à présent de combiner les différents indices contextuels que nous avons manipulés jusqu'à présent. Comme nous l'avons expliqué en introduction de ce chapitre, nous allons combiner quatre indices contextuels :

- La catégorisation thématique (section 6.1) : c'est un indice contextuel spécifique à l'application. Cet indice s'est révélé être un outil exploratoire utile pour faire émerger la diversité portée par une requête.
- La temporalité (section 5.4.4) : l'étude de ces requêtes a fait émerger deux types de profils : les requêtes ponctuelles et les requêtes durables. Si jusqu'à présent, l'ambiguïté des requêtes a été envisagée en synchronie, il nous paraît pertinent de mettre en lumière un facteur supplémentaire de variation : la dimension diachronique. En effet, nous avons pu observer que certaines requêtes changent de catégorisation thématique selon les périodes temporelles.
- L'extension des requêtes : cet indice est utilisé en complément. Il permet d'apporter un contexte utilisateur manquant.
- Les cooccurrences : ce dernier indice est également complémentaire aux autres indices, particulièrement à la catégorisation thématique. Ce sont deux indices qui permettent d'accéder aux informations contextuelles situées dans les documents.

Nous avons réuni les différents indices dans le tableau 8.11. Ce tableau mentionne les types de contexte (sections 3.2.2 et 4.3) et la source de ces indices (sections 4.3 et 5.1.1). Nous allons tenter de faire émerger des hypothèses à

partir d'analyses de cas de requêtes présentes dans notre corpus. Les requêtes que nous allons étudier sont issues du corpus *2424reqFréquentes.*

Indices	Types de contexte	Source
Catégorisation thématique	- Contexte de l'information	Documents, Méta-données
Temporalité	- Contexte de l'information - Contexte utilisateur	Historique de recherche (log)
Requêtes étendues	- Contexte utilisateur	Requêtes
Cooccurrence	- Contexte de l'information	Documents

TABLEAU 8.11 : Caractéristiques des indices combinés

8.3.1 Analyse d'une requête pluricatégorisée : *sarkozy*

La requête *sarkozy* est un NPP. Cette requête a été présente parmi les requêtes populaires pendant 5 mois. La combinaison des indices contextuels révèle une requête complexe, cumulant plusieurs types d'ambiguïté. Ces indices sont présentés dans le tableau 8.12 sous la forme d'une carte synthétique [1].

1. Les cooccurrents sont calculés dans une fenêtre de 3 mots à gauche et à droite et ils ont une fréquence supérieure à 3. Les collocats identiques sont marqués en gras.

Temps	Catégorisation	Cooccurrents	Requêtes étendues
Mai	ÉCO/POL/SOC	élèves, tord, politise, obsèques, enquiert	*jean sarkozy* (6) - *nicolas sarkozy* (4) - *sarkozy an iii* (2) - *sarkozy bouteille* (2)
Juin	INT/POL/SOC	égratignant, pousuivre, financiers, exalté, **devancerait**	*nicolas sarkozy* (14) - *jean sarkozy* (6) - *clash sarkozyaubry* (1) - *sarkozy mitterrand* (1)
Juil	INT/POL/SOC	parole, Intervention, gouvernementNicolas, expulsons, exhorte	/
Aout	INT/POL/SOC	regagne, **devancerait**, confiance, réprobation, play	/
Oct	ÉCO/POL/SOC	vibrion, tait, rapprochement, portée, popularité	*nicolas sarkozy* (9) - *jean sarkozy* (4) - *lascaux et sarkozy* (3) - *carla bruni sarkozy* (2)

TABLEAU 8.12 : Carte des indices contextuels de la requête *sarkozy*

La requête *sarkozy* a un comportement intéressant vis-à-vis de la catégorisation comme on peut le voir dans le tableau 8.12. Elle a une tendance forte et constante à la pluri-catégorisation. En effet, en mai et octobre la requête a pour catégorisation trois thématiques qui sont ÉCONOMIE, POLITIQUE et SOCIÉTÉ. Les mois de juin, juillet et août sont catégorisés en INTERNATIONAL, POLITIQUE et SOCIÉTÉ. Nous allons donc nous intéresser aux contextes d'apparition de cette requête grâce à l'apport des cooccurrences.

Les collocats les plus proches statistiquement du mot *sarkozy* sont réunis dans le tableau 8.12. Les collocats présentent une variation en diachronie. On observe qu'une grande part des collocats sont des verbes (*tord, politise, enquiert, Expulsons, exhorte, regagne*). On relève également des collocats qui dépendent de l'actualité comme en mai avec *élèves* et *obsèques* ou *financiers* en juin. En juillet, les collocats révèlent que le mot *sarkozy* est fortement lié à la prise de parole du président de la république : *intervention, exhorte, parole* (pour prise

de parole)[2]. Enfin, en octobre, deux collocats, *vibrion*[3] et *popularité*, renvoient à la popularité de Nicolas Sarkozy, en berne. On note également des erreurs héritées des corpus telles que la présence du mot « pousuivre » parmi les collocats, montrant l'inadéquation de la mesure d'information mutuelle et les limites de notre extraction de cooccurrences.

Les versions étendues de cette requête montrent un autre type de variation comme on le voit dans le tableau 8.12 où sont présentées les requêtes étendues les plus fréquentes. Les requêtes étendues comportent différents noms de personnes comme Jean Sarkozy ou Carla Bruni-Sarkozy. Ces requêtes étendues montrent la présence d'homonymie. Elles sont également nombreuses à comporter des éléments de spécifications géographiques comme *lascaux et sarkozy*, ou à spécifier des relations avec d'autres personnes comme *clash sarkozyaubry* ou *sarkozy mitterrand.*

L'apport des indices contextuels réside dans le fait qu'ils permettent de révéler au moins deux types de problème touchant cette requête. En effet, la requête *sarkozy* est ambiguë lexicalement, à cause d'homonymies, indiquées par les requêtes étendues. Les catégories thématiques, indice contextuel provenant de la base documentaire, signalent aussi que cette requête peut être vague ou large, sous l'effet du manque de contexte.

8.3.2 Analyse d'une requête fortement étendue : *grève*

Nous proposons d'analyser le cas de la requête *grève* car elle est une illustration d'une requête large ou vague mais qui est fortement étendue. C'est une requête qui a été parmi les plus populaires de septembre à décembre 2010 (4 mois). Les indices contextuels sont réunis dans le tableau 8.13 sous la forme d'une carte synthétique. La combinaison de ces indices permet de réduire l'indétermination portée par cette requête.

La catégorisation thématique de la requête *grève* varie dans le temps, comme on peut le voir dans le tableau 8.13. En septembre et décembre, la requête est rattachée à trois catégories, dont deux catégories communes, INTERNATIONAL et SOCIÉTÉ. Les deux autres périodes restantes sont moins propices à une catégorisation multiple. En effet, la requête est même mono-catégorisée en novembre 2010. La catégorisation signale donc que cette requête semble sensible aux variations de l'actualité.

2. Le président Sarkozy a prononcé le « discours de Grenoble » en juillet 2010 donnant lieu à un grand nombre de commentaires dans la presse.
3. Emploi familier désignant une personne agitée.

Temps	Catégorisation	Cooccurrents	Requêtes étendues
Sept	INT / SOC / POL	postiers, notice, NOTAM, **illimitée**, Oups	/
Oct	ÉCO / SOC	déchets, rectangle, reconduisent, **Préavis**, Onzième	***greve rer* (12 179)** - *greve sncf* (43) - *greve air france* (33) - *preavis de greve* (24)
Nov	ÉCO	**Préavis/préavis**, piquet, métros , Portugal, **générale**	/
Déc	INT / SOC / ÉCO	**illimitée,** reconductible , **préavis,** contrôleurs, **générale**	*greve rer* (16) - *total greve* (11) - *greve bordeaux* (3) - *greve montpellier* (3)

TABLEAU 8.13 : Carte des indices contextuels de la requête *grève*

Nous cherchons dans un deuxième temps ces premiers éléments de variation mis en évidence par la catégorisation thématique grâce à l'analyse des cooccurrents fréquents de la requête dans les documents de la base textuelle. Les collocats révélés par le logiciel Antconc sont présentés dans le tableau 8.13. Nous constatons que, malgré une variation en diachronie, les mêmes collocats se retrouvent dans les différents corpus. En effet, on voit que les collocats du mot *grève* sont relativement semblables : *illimitée, préavis, générale*. Ces collocats forment des unités lexicales spécifiant le mot *grève* comme « grève illimitée », « préavis de grève » ou encore « grève générale ». Du point de vue diachronique, nous retrouvons des collocats qui spécifient le type de personnes qui font grève comme *postiers* ou *contrôleurs*, alors que d'autres collocats spécifient le lieu de la grève (*portugual*) ou le nombre de jours de grève (*onzième, rectangle*). Cette analyse nous montre que le mot *grève* est nécessairement spécifié en contexte et que certaines associations sont privilégiées.

En observant les requêtes étendues de la requête *grève*, on voit que celle-ci est étendue de manière massive (ratio de 0,47) au mois d'octobre 2010 en *grève rer* (tableau 8.13). Cette requête étendue spécifie ainsi l'information de *grève* vers une cible particulière qu'est le RER parisien. Les autres requêtes étendues concernent également des grèves dans les transports comme *grève air france* ou *grève sncf*. Les versions étendues de la requête *grève* au mois d'octobre comportent aussi des termes comme *mouvement de grève ou préavis de*

grève. Or, au mois de décembre, ces requêtes étendues changent et leurs fréquences sont moins élevées (le ratio est seulement de 0,03). Le cas de la requête *grève rer* est le plus marquant, puisqu'il n'y a eu que 16 occurrences de cette requête dans le log de requêtes du mois de décembre. La plupart des requêtes étendues spécifient géographiquement *grève* comme dans *grève montpellier* ou *grève bordeaux* ce qui montre la nécessité de spécification de cette requête. Le changement observé entre le mois d'octobre et de décembre montre la grande variation de l'actualité ainsi que la grande volatilité des centres d'intérêts des utilisateurs.

La combinaison des indices contextuels montre que cette requête est vague si elle n'est pas spécifiée. Cette indétermination est accentuée par la présence d'un mouvement social contre la réforme des retraites qui a généré un certain nombre de grèves en France au mois d'octobre 2010. Cela se traduit au niveau des requêtes des utilisateurs et des documents de la base. Ce cas montre également que des variations temporelles peuvent toucher une requête.

8.3.3 Analyse d'une requête ponctuelle : *france2*

La requête *france2* a une durée de vie assez limitée (3 mois) d'octobre à décembre 2010. Nous avons consigné dans le tableau 8.14 l'ensemble des indices disponibles. Cette requête n'est pas ambiguë d'un point de vue linguistique ou vague, c'est un changement de fonction de la requête au cours d'une période temporelle qui provoque une ambiguïté.

Il s'avère que cette requête n'a été pluri-catégorisée qu'une seule fois en octobre 2010. En décembre 2010, elle n'a pas été catégorisée, aucune occurrence n'ayant été trouvée dans le texte des documents. Elle correspond toutefois à une source d'information de l'application 2424actu. Or la catégorisation de cette requête en octobre et novembre met en évidence un emploi de cette EN, non pas comme source ou média d'information, mais comme lieu et acteur d'un évènement, en l'occurrence cela fait référence à ce que l'on a appelé l'affaire Guerlain, précisément au journal télévisé de 13h.

Le calcul des cooccurrents de *france2* vient conforter ces premières observations. En effet, il n'existe pas de collocats utilisables pour la forme de cette requête. Les collocats extraits sont des métadonnées, montrant le statut de source d'actualité de *france2*. Il s'avère que l'on retrouve ces informations parmi les versions étendues de cette requête en octobre 2010 : *13h jt france2,13h jt france219 octobr, 13h jt france219 octobre* (tableau 8.14). En décembre 2010, les versions étendues de la requête *france2* indiquent que certaines personnes cherchent toujours des informations à propos du fait-divers présenté ci-dessus,

Temps	Catégorisation	Cooccurrents	Requêtes étendues (fréquence = 1)
Oct	POL/SOC	NULL, int, clt, DATE, SOURCE	*france2 210ct - france2 jt - france2 lyon - france2 meteo - journal france2 12 septembre 2009 - les actualites lundi soir france2 - 13h jt france2 - 13h jt france219 octobr - 13h jt france219 octobre - xavier de franssu france2 -*
Nov	POL	NULL, int, soc, Canal, DATE	/
Déc	/	/	*france23 - france2 homme 85 ans se suicide maison de retraite - journal france2 - jt 13h france2, langres jt 13h france2 - france2 hopital espoir - france2 infos*

TABLEAU 8.14 : Carte des indices contextuels de la requête *france2*

mais c'est l'emploi en tant que source qui est recherchée par le reste des utilisateurs.

Nous pouvons supposer que cette requête est navigationnelle puisqu'aucune information sur *france2* n'est présente dans la base documentaire. Puis elle a été ponctuellement informationnelle lors de l'apparition d'un fait-divers en octobre 2010 comme on le voit dans les requêtes étendues. Ce changement de type de requête créé une ambiguïté de type homonymique entre une source d'information et une institution.

8.3.4 Analyse d'une requête durable et pluracatégorisée : *haïti*

Nous terminons cette analyse de cas par l'étude de la requête *haïti*. Cette requête est présente tout au long de notre corpus de requêtes populaires (cf. section 5.4.4). C'est cette longévité qui est intéressante, signe d'une possible mobilité de cette requête. Pour l'étudier, nous mobilisons l'ensemble des indices contextuels disponibles comme pour les autres requêtes (tableau 8.15).

La requête *haïti* a un comportement intéressant vis-à-vis de la catégorisation thématique. Cette requête est la plupart du temps catégorisée en INTERNATIONAL (au mois de juin, août, octobre et novembre). Malgré une tendance forte à la mono-catégorisation, on observe plusieurs changements. Au mois de juin, la requête est catégorisée en CULTURES et en SPORT. Au mois de décembre, la requête est catégorisée à la fois en SOCIÉTÉ et en INTERNATIONAL. Nous savons également que la thématique INTERNATIONAL est très englobante (cf. chapitre 6). On peut donc supposer que cette requête a une capacité de variation forte.

Dans un deuxième temps, nous observons si cette variation mise en évidence par la catégorisation thématique se retrouve dans les documents de la base textuelle, pour cela nous réalisons une analyse des cooccurrents fréquents et fortement liés au terme *haïti* sur plusieurs périodes temporelles (mai, août, octobre, novembre, décembre). L'hypothèse est que la variation se traduit par la présence de cooccurrents différents selon la période temporelle.

Dans le tableau 8.15, sont présentés les cooccurrents les plus fréquents de *haïti* à différentes époques temporelles. Nous constatons qu'ils sont effectivement très différents. Nous discernons plusieurs évènements importants comme l'annonce des élections au mois d'août 2010 : *élection, invalidée, rappeur* (candidature du rappeur Wyclef Jean). Au mois d'octobre, c'est l'épidémie de choléra qui apparaît (*éradiqué, ralentie, kits, choléra*) suivie de l'ouragan Tomas au mois de novembre. Le seul cooccurrent de *haïti* présent sur plusieurs mois est Minustah qui désigne la mission des Nations Unies pour la stabilisation du pays,

Temps	Catégorisation	Cooccurrents	Requêtes étendues
Mai	CLT/INT/SOC	Crowe, Etats, unis, Forte, Russell	*haïti adoption* (26) - *haïti séisme* (2) - *adoption haïti* (1) - *haïti cacao* (1)
Juin	INT/SOC/ÉCO	/	*haïti adoption* (26) - *haïti actu* (2) - *images du séisme en haïti* (2)
Juil	CLT/SPR	/	
Aout	INT	quittait, invalidée, Minustah, élection, rappeur	*haïti adoption* (8) - *haïti 17* (3) - *haïti aujourd'hui* (2) - *haïti séisme* (2) - *tremblement de terre à haïti* (2) -
Sept	INT	/	*haïti adoption* (4) - *radio d'haïti* (2) - *haïti reconstruction* (1)
Oct	INT	Éradiqué, ralentie, kits, adoptions, choléra	*haïti adoption* (11) - *élection haïti* (4) - *adoption haïti* (2)
Nov	INT	Tomas, Ouragan, Casques, Minustah, évêques	*haïti adoption* (16) - *haïti choléra* (9) - *bilan d'octobre à aujourd'hui haïti choléra* (4)
Déc	INT/SOC	adoptés, passeraient, Minustah, secouent, recomptage	*haïti adoption* (68) - *haïti élection* (11) - *haïti actualités 11 12 2010* (3) - *haïti choléra* (3) - *haïti jude célestin* (3) - *élections haïti* (2)

TABLEAU 8.15 : Carte des indices contextuels de la requête *haïti*

qui constitue de fait un arrière-plan stable. On détecte donc une variation des évènements associés au terme *haïti*, induite par cette actualité mouvementée.

L'analyse des versions étendues de la requête *haiti* tout au long du corpus (8.15) vient conforter les observations faites sur les cooccurrents de *haïti* en contexte. En effet, les utilisateurs produisent des versions étendues qui permettent d'identifier des thèmes associés à l'actualité d'Haïti. On retrouve en particulier les thématiques des élections et de l'adoption. Mais l'absence de requêtes étendues peut aussi être un indice intéressant. En effet, au mois d'octobre, alors que l'épidemie de choléra touche l'île comme nous pouvons le voir dans les cooccurrences, le mot *choléra* n'apparait pas dans les requêtes étendues. Il apparaît seulement en novembre. On suppose donc que la requête *haïti* donnait accès en priorité aux informations sur cette épidémie au mois d'octobre.

L'analyse que nous avons effectuée sur la requête *haiti* fait intervenir trois faisceaux d'information : la catégorisation thématique, la distribution de la requête dans les textes et les versions étendues de cette requête. La catégorisation nous a surtout montré que les thématiques liées à cette requête pouvaient évoluer, malgré un ancrage fort dans la catégorie INTERNATIONAL. Nous savons que cette thématique est difficile à interpréter et qu'elle cache potentiellement une diversité plus grande. L'analyse des cooccurrents de la requête *haïti* dans les documents et l'analyse des requêtes étendues ont effectivement confirmé cette diversité.

En effet, on peut observer au moins deux emplois possibles du mot *haïti* : comme référence au tremblement de terre ou dans un sens locatif. Haïti semble manifester une polyvalence de base (le lieu, le pays et les habitants) et des variations contextuelles propres à l'actualité. Ainsi lorsque le pays Haïti a été touché par une épidémie de choléra, les cooccurrents de mot *Haïti* dans les documents ont changé. La requête renvoie à des informations différentes, touchée par une variation contextuelle.

8.4 Conclusion

L'association des indices contextuels montre que ces indices sont complexes et sont parfois en décalage. Le contexte utilisateur ne donne pas forcément les mêmes informations que les documents de la base documentaire. Ces décalages sont tout de même intéressants, ils permettent de révéler le fonctionnement de certaines requêtes. D'autre part, certains indices ne sont pas automatisables en l'état, comme les cooccurrences.

L'indice de temporalité a un rôle déterminant, il permet de mettre en évidence des comportements sur le long terme comme la polyvalence de *haïti* mais aussi des comportements ponctuels comme une variation de type évènementiel. L'indice d'extension d'une requête n'est pas à utiliser de manière isolée concernant une suspicion d'ambiguïté. Mais il permet de montrer que les utilisateurs ont conscience des différentes interprétations que peuvent prendre certaines requêtes.

Le croisement des indices contextuels révèle l'aspect fortement volatile de l'actualité. En effet, cette volatilité a un impact sur les requêtes et sur l'ambiguïté comme ont pu le montrer les requêtes *grève* ou *haïti*. Nous avons choisi de proposer une solution qui tenait compte des spécificités de l'application et donc des variations imposées par l'actualité, en l'occurrence la catégorisation thématique, solution que nous avons ensuite complétée de manière exploratoire avec la présence des requêtes étendues. Pour prolonger cette démarche, nous pensons qu'il serait intéressant de procéder à un co-monitoring des requêtes et des sujets présents dans l'actualité, et d'identifier les cas de « ruptures ». Une telle démarche permettrait de pouvoir systématiser les observations que nous avons développées dans ce chapitre.

Conclusion et perspectives

Conclusion

Notre principale réalisation a consisté à mettre en place un dispositif pour étudier l'ambiguïté des requêtes. Ce dispositif repose sur l'hypothèse que les indices contextuels en RI sont des éléments déterminants pour révéler et comprendre l'ambiguïté qui touche les requêtes. La démarche que nous avons voulu suivre a été complexe à mettre en place. Nous avons dû nous adapter aux problèmes spécifiques imposés par un contexte industriel basé sur la confidentialité des données : moteur de recherche non accessible, anonymisation des données utilisateurs, format des métadonnées en perpétuelle évolution, données issues de l'agrégateur non diffusables.

Ce travail nous a permis de constater que l'ambiguïté des requêtes n'est pas similaire à l'ambiguïté telle que définie en linguistique. En effet, l'ambiguïté des requêtes est une notion plus large recouvrant également les requêtes vagues ou excessivement génériques. L'ambiguïté des requêtes est une manifestation qui ne peut se définir que par rapport à un référentiel lexicographique. En cela, l'ambiguïté se diagnostique au regard des résultats ramenés par le moteur de recherche et peut également être appréciée au regard de la capacité de l'utilisateur à interpréter les résultats de recherche. C'est la raison pour laquelle nous avons privilégié une approche cherchant à repérer la diversité présente dans les résultats de recherche ramenés par une requête et intégrant le recours au jugement des utilisateurs. Cette démarche nous a permis d'identifier divers comportements comme les aspects évènementiels des noms propres ou les requêtes larges qui ne correspondent pas à de l'ambiguïté classique. Ce sont des comportements linguistiques particulièrement intéressants de requêtes composées de noms propres parce que ce sont des variations contextuelles propres à l'actualité qui ne sont pas connues à l'avance.

La méthode de catégorisation thématique est une manière d'accéder à l'ambiguïté portée par une requête. Nous avons montré que cette ambiguïté est

différente de celle repérée par une ressource encyclopédique telle que Wikipédia, ressource lexicographique. Cette différence est due au fait que la catégorisation thématique capte mieux la réalité des emplois du corpus d'actualité. L'utilisation des catégories thématiques nous a permis d'amorcer le typage de l'ambiguïté des requêtes. Cependant, cette typologie demande à être éprouvée à la fois sur une plus grande quantité de requêtes, mais également sur des requêtes provenant d'autres moteurs de recherche. Nous avons donc évalué la catégorisation en mettant en place un test utilisateur.

Le premier test utilisateur réalisé a donné lieu à la production de labels commentant les regroupements thématiques de résultats de recherche. L'emploi du moteur de recherche a permis d'améliorer et de tester le dispositif de catégorisation. Nous avons dû ensuite trouver un moyen adapté pour évaluer les labels produits, particulièrement l'homogénéité des labels indiquant une compréhension du dispositif. Ce test utilisateur nous a permis de comprendre que notre dispositif est adapté pour révéler la diversité de plusieurs de nos requêtes testées. Il nous a aussi permis de voir que certaines catégories thématiques sont trop généralistes et que cela produit des regroupements de documents moins pertinents pour les utilisateurs.

Le deuxième test utilisateur, plus qualitatif, a permis de comparer la méthode de catégorisation automatique des résultats à une catégorisation humaine. Nous avons ainsi démontré que les deux modes de catégorisation se recoupaient particulièrement sur les résultats de certaines requêtes que nous avons clairement identifiées. Toutefois, les limites de ce travail réside dans la possibilité d'une reproductibilité. En effet, les sujets ont déclaré que leurs classements seraient certainement différents s'ils devaient les refaire. Il faudrait donc tester à une plus grande échelle cette tâche.

D'autre part, nous avons testé et étudié à plus ou moins grande échelle un certain nombre d'indices contextuels dans le but d'observer le comportement des requêtes de manière globale. Pour cela, nous avons analysé finement les requêtes et les documents dans lesquelles elles apparaissaient. Cela a demandé la réalisation d'un véritable travail d'expertise linguistique sur les requêtes des utilisateurs, mais aussi sur l'actualité.

Nous avons utilisé un faisceau d'indices contextuels (longévité des requêtes, proportion de versions étendues, diversité des contextes dans la base documentaire). Ce faisceau nous a fourni des informations sur le comportement d'une requête. Cependant, lorsque nous avons couplé l'ensemble des indices dont nous disposions, nous avons constaté que ces indices apportaient des informations différentes. Ce sont parfois ces informations contradictoires qui permettent de mettre en évidence certains fonctionnements comme des va-

riations de type navigationnel / informationnel. Mais cela apporte également une difficulté supplémentaire lorsque l'on envisage une automatisation.

Tout au long de nos expérimentations, nous nous sommes appuyée sur un typage sémantique des requêtes. Quelles conclusions en tirer ? Les requêtes comportant des noms communs sont composées de mots majoritairement non autonomes du point de vue référentiel, ce qui crée une ambiguïté dans de nombreux cas identifiés. Les noms propres de personne se sont avérés être des requêtes difficiles à traiter, les indices contextuels donnent en effet des informations contradictoires à leur sujet. Les extensions de requêtes sont particulièrement utiles sur ces types de requêtes, éclairant sur l'intention des utilisateurs. Les noms propres de lieux, qui sont majoritairement des noms de pays, méritent d'être étudiés de manière plus approfondie en diachronie. Ils forment des requêtes fortement malléables, réceptives aux variations de l'actualité.

Nous avons été confrontée à travers cette étude au décalage de perception de l'actualité entre les utilisateurs du moteur et de l'agrégateur et les médias d'actualité. En effet, les sujets les plus traités par les sources d'informations ne sont pas forcément les plus recherchés par les utilisateurs. En général, les sujets fortement traités sont facilement à disposition et par conséquent il n'est pas nécessaire de les rechercher. Nous avons également fait face à la prédominance des faits-divers parmi les évènements recherchés autant dans les requêtes que dans les sujets traités par les médias. Les faits-divers donnent eux aussi lieu à des requêtes ambigues et vagues. On peut en effet supposer que la forte présence du fait-divers dans les médias fait « oublier » à l'utilisateur que ce n'est pas le seul sujet traité. L'exemple de la requête *laetitia* le montre bien (cf. chapitre 7).

Perspectives

Nos travaux ont apporté des réponses à plusieurs de nos interrogations initiales, mais des points restent encore à explorer.

En premier lieu, des améliorations peuvent être apportées au dispositif de catégorisation. Une amélioration de l'étiquetage des documents serait appréciable, nous avons pu en effet constater les multiples erreurs générées par la phase d'étiquetage des documents. En effet, lors de l'évaluation, nous avons observé que certaines catégories sont trop larges et produisent des regroupements de documents de faible qualité. Cependant, l'étiquetage expert conserve sa pertinence. Les problèmes de qualité de la catégorisation sont les consé-

quences de choix techniques opérés par la plate-forme 2424actu, en aval de l'étiquetage expert.

Nous pourrions envisager d'améliorer le dispositif de catégorisation des résultats en utilisant un travail éditorial sur les résultats d'un moteur d'actualités. D'une part, il faudrait travailler sur un mode de présentation des résultats qui prend en compte le contenu éditorial des sources. Et d'autre part, il faudrait mieux sélectionner les sources d'actualités. En effet, même si dans le cas de 2424actu, les médias d'actualité utilisés comme source ont été sélectionnés, la qualité de l'information est relativement hétérogène.

Les différentes observations réalisées sur les requêtes et leur capacité à être ambiguës nous laisse supposer qu'il serait intéressant de tester des dispositifs en fonction du type sémantique de la requête. Par exemple, nous pourrions tester l'apport d'un dispositif de personnalisation sur les requêtes de type NC ou de l'expansion de requêtes sur les NPL ou NPP. Il reste également à explorer une comparaison avec une approche de clustering de documents, que nous avons écartée dans ce travail. Même si nous l'avons envisagé, il nécessiterait la mise en place d'un test utilisateur de plus grande envergure, en terme de travail de développement mais aussi en temps de passation avec le recrutement d'au moins deux groupes de 20 utilisateurs. Enfin, il faudra veiller à ce que les résultats du clustering aient une granularité similaire, ce qui n'est pas évident sur les documents d'actualité.

Les indices contextuels que nous avons commencé à explorer demandent un traitement plus systématique que l'analyse qualitative réalisée dans le cadre de cette thèse. Il serait intéressant de travailler sur un dégroupage sémantique des cooccurrents, et par conséquent d'appliquer un nouveau traitement au corpus de documents. Ce dégroupage permettrait de faire émerger automatiquement des groupes de collocats marquant des emplois particuliers de l'acception étudiée. L'indice concernant la capacité des requêtes à être étendues (cf. chapitre 8) mériterait d'être remplacé par des reformulations car il serait alors possible de se focaliser sur un utilisateur en particulier. Cela permettrait aussi d'accéder aux parcours de recherche des utilisateurs et donc de détecter des reformulations qui n'ont aucune forme en commun.

Nous avons également vu que l'étude en diachronie et en synchronie apportait des réponses différentes. D'un point de vue diachronique, les requêtes semblent manifester plus d'ambiguïtés. Ceci peut avoir un impact important lorsque la base documentaire donne accès à de l'actualité sur plusieurs années comme le fait par exemple l'INA ou le faisait Google Actualités.

Mais ces variations constatées peuvent offrir une piste applicative découlant

de cette recherche sur l'ambiguïté pourrait être un outil de veille et d'analyse de l'actualité. En effet, nous avons montré que les indices contextuels détectent les nombreuses variations de l'actualité. Cela pourrait permettre d'identifier des ruptures intéressantes dans le flux de l'actualité, élément utile pour la prédiction de l'émergence d'un nouveau sujet d'actualité. L'étude des requêtes est également un moyen de prévoir l'apparition de nouvelles tendances chez les utilisateurs avant que les médias s'emparent véritablement de la question. Par exemple, la requête *tunisie* est apparue dès le début du mois de décembre, bien avant que les médias parlent des débuts de la révolution. En effet, un moteur d'actualités peut être un recours pour un utilisateur qui cherche une information « rare » et non traitée par les médias.

Bibliographie

Clémentine ADAM, Cécile FABRE et Ludovic TANGUY : Etude des relations sémantiques dans les reformulations de requêtes sous la loupe de l'analyse distributionnelle. *In Actes de SemDis 2013 : Enjeux actuels de la sémantique distributionnelle*, pages 140–153, Les Sables d'Olonne, France, 2013.

Eneko AGIRRE, Olatz ANSA, Eduard H. HOVY et David MARTINEZ : Enriching wordnet concepts with topic signatures. *In Proceedings of the NAACL Workshop on WordNet and Other lexical Ressources : Applications, Extensions and Customizations*, pages 23–28, 2001.

Thomas E. AHLSWEDE : Word sense disambiguation by human informants. *In Proceedings of the Sixth Midwest Artificial Intelligence and Cognitive Society Conference*, pages 73–78, 1995.

Thomas E. AHLSWEDE et David LORAND : The ambiguity questionnaire : A study of lexical disambiguation by human informants. *In Proceedings of the Fifth Midwest Artificial Intelligence and Cognitive Society Conference*, pages 21–25, 1993.

James ALLAN, Jay ASLAM, Nicholas BELKIN, Chris BUCKLEY, Jamie CALLAN, Bruce CROFT, Sue DUMAIS, Norbert FUHR, Donna HARMAN, David J. HARPER, Djoerd HIEMSTRA, Thomas HOFMANN, Eduard HOVY, Wessel KRAAIJ, John LAFFERTY, Victor LAVRENKO, David LEWIS, Liz LIDDY, R. MANMATHA, Andrew MCCALLUM, Jay PONTE, John PRAGER, Dragomir RADEV, Philip RESNIK, Stephen ROBERTSON, Roni ROSENFELD, Salim ROUKOS, Mark SANDERSON, Rich SCHWARTZ, Amit SINGHAL, Alan SMEATON, Howard TURTLE, Ellen VOORHEES, Ralph WEISCHEDEL, Jinxi XU et ChengXiang ZHAI : Challenges in information retrieval and language modeling : report of a workshop held at the center for intelligent information retrieval, university of massachusetts amherst, september 2002. *SIGIR Forum*, 37(1):31–47, 2003.

James ALLAN et Hema RAGHAVAN : Using part-of-speech patterns to reduce query ambiguity. *In Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '02, pages 307–314. ACM, 2002.

L. ANTHONY : Antconc (version 3.2.2) [computer software]. Rapport technique, 2011. URL http://www.antlab.sci.waseda.ac.jp/.

Laurent AUDIBERT : *Outils d'exploration de corpus et désambiguïsation lexicale automatique*. Thèse de doctorat, Université de Provence - Aix-Marseille I, 2003.

Cory BARR, Rosie JONES et Moira REGELSON : The linguistic structure of english web-search queries. *In Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08, pages 1021–1030. Association for Computational Linguistics, 2008.

N. J. BELKIN, R. N. ODDY et H. M. BROOKS : ASK for information retrieval : part I. Background and theory. *Documentation*, 38(2):61–71, 1982a.

N. J. BELKIN, R. N. ODDY et H. M. BROOKS : ASK for information retrieval : part II. results of a design study. *Documentation*, 38(3):145–164, 1982b.

Andrea BERNARDINI, Claudio CARPINETO et Massimiliano D'AMICO : Full-subtopic retrieval with keyphrase-based search results clustering. *In Proceedings of the 2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology*, volume 1 de *WI-IAT '09*, pages 206–213, Washington, DC, USA, 2009. IEEE Computer Society.

Frédérique BIVILLE : *Polysémie et noms propres*, pages 37–50. PUPS, 2005.

Robert BOURE et Nikos SYMRNAIOS : L'infomédiation de l'information en ligne. Le cas des filiales françaises de Google et Yahoo. pages 43–55, 2006.

Mokrane BOUZEGHOUB et Dimitre KOSTADINOV : Personnalisation de l'information : aperçu de l'état de l'art et définition d'un modèle flexible de profils. *In CORIA'05*, pages 201–218, 2005.

Sergey BRIN et Lawrence PAGE : The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems*, 30:107–117, 1998.

Andrei BRODER : A taxonomy of web search. *SIGIR Forum*, 36(2):3–10, septembre 2002.

Paul BUITELAAR, Bernardo MAGNINI, Carlo STRAPPARAVA et Piek VOSSEN : *Domain-specific WSD*, pages 275–252. Springer, 2006.

Claudio CARPINETO, Stanislaw OSIŃSKI, Giovanni ROMANO et Dawid WEISS : A survey of web clustering engines. *ACM Computing Surveys (CSUR)*, 41(3):17–38, 2009.

Jean-Marie CHARON et Patrich Le FLOCH : *La presse en ligne*. La Découverte, 2011.

Hao CHEN et Susan DUMAIS : Bringing order to the web : automatically categorizing search results. *In Proceedings of the SIGCHI conference on Human factors in computing systems*, CHI '00, pages 145–152, New York, NY, USA, 2000. ACM. ISBN 1-58113-216-6.

Harr CHEN et David R. KARGER : Less is more : probabilistic models for retrieving fewer relevant documents. *In Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '06, pages 429–436, New York, NY, USA, 2006. ACM.

Max CHEVALIER : *Usagers et Recherche d'information*. Habilitation à diriger des recherches, Université Paul Sabatier, 2011.

Paul Alexandru CHIRITA, Wolfgang NEJDL, Raluca PAIU et Christian KOHLSCHÜTTER : Using odp metadata to personalize search. *In Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '05, pages 178–185. ACM, 2005.

Charles LA CLARKE, Maheedhar KOLLA et Olga VECHTOMOVA : An effectiveness measure for ambiguous and underspecified queries. *In Advances in Information Retrieval Theory*, pages 188–199. Springer, 2009.

Paul CLOUGH, Mark SANDERSON, Murad ABOUAMMOH, Sergio NAVARRO et Monica Lestari PARAMITA : Multiple approaches to analysing query diversity. *In SIGIR*, pages 734–735, 2009.

Jacob COHEN : A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46, 1960.

Thomas M COVER et Joy A THOMAS : *Elements of information theory*. Wiley-interscience, 2nd edition édition, 2012.

Daniel CRABTREE, Xiaoying GAO et Peter ANDREAE : Improving web clustering by cluster selection. *In Proceedings of the 2005 IEEE/WIC/ACM International Conference on Web Intelligence*, WI '05, pages 172–178, Washington, DC, USA, 2005. IEEE Computer Society.

W. Bruce CROFT : Knowledge-based and statistical approaches to text retrieval. *IEEE Expert : Intelligent Systems and Their Applications*, 8(2):8–12, avril 1993.

William CROFT et Alan CRUSE : *Cognitive Linguistics*. Cambridge University Press, 2004.

Steve CRONEN-TOWNSEND et W. Bruce CROFT : Quantifying query ambiguity. *In Proceedings of the second international conference on Human Language Technology Research*, HLT '02, pages 104–109. Morgan Kaufmann Publishers Inc., 2002.

Alan CRUSE : *Lexical Semantics*. Cambridge University Press, Cambridge, UK, 1986.

Kareem DARWISH et Douglas W. OARD : Probabilistic structured query methods. *In Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, SIGIR '03, pages 338–344, New York, NY, USA, 2003. ACM.

Fernando DIAZ et Rosie JONES : Using temporal profiles of queries for precision prediction. *In Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '04, pages 18–24, New York, NY, USA, 2004. ACM.

Zhicheng DOU, Ruihua SONG et Ji-Rong WEN : A large-scale evaluation and analysis of personalized search strategies. *In Proceedings of the 16th international conference on World Wide Web*, WWW '07, pages 581–590, New York, NY, USA, 2007. ACM.

Susan DUMAIS, Edward CUTRELL et Hao CHEN : Optimizing search by showing results in context. *In Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 277–284. ACM, 2001.

Maud EHRMANN : *Les Entités Nommées, de la linguistique au TAL : statut théorique et méthodes de désambiguïsation.* Thèse de doctorat, Paris VII, 2008.

Christian FELLBAUM, éditeur. *WordNet : An Electronic Lexical Database (ISBN : 0-262-06197-X)*. MIT Press, first édition, 1998.

W. Nelson FRANCIS et Henry KUCERA : *Frequency analysis of English usage : Lexicon and grammar*. Boston : Houghton Mifflin Company, 1982.

Catherine FUCHS : L'hétérogénéité interprétative. *Documenti di Lavoro e Pre-Publicazioni*, 180-182:1–15, 1989.

Catherine FUCHS : Ambiguïté et ambivalence : le discret et le continu. *Les cachiers du CRIAR*, pages 9–23, 1994.

Catherine FUCHS : *Les ambiguïtés du français*. Ophrys, 1996.

William A. GALE, Kenneth W. CHURCH et David YAROWSKY : One sense per discourse. *In Proceedings of the DARPA Speech and Natural Language Workshop*, pages 233–237, 1992.

William A. GALE, Kenneth W. CHURCH et David YAROWSKY : A method for disambiguating word senses in a large corpus. *Computers and the Humanities*, 26:415–439, 1993.

Qingqing GAN, Josh ATTENBERG, Alexander MARKOWETZ et Torsten SUEL : Analysis of geographic queries in a search engine log. *In Proceedings of the first international workshop on Location and the web*, LOCWEB '08, pages 49–56, New York, NY, USA, 2008. ACM. ISBN 978-1-60558-160-6.

Marie-Noëlle GARY-PRIEUR : *L'individu pluriel. Les noms propres et le nombre.* CNRS Éditions, 2001.

Alfio GLIOZZO et Bernardo MAGNINI : Unsupervised domain relevance estimation for word sense disambiguation. *In In Proceedings of International Conference on Empirical Methods in Natural Language Processing (EMNLP'04*, pages 380–387, 2004.

Julio GONZALO, Felisa VERDEJO, Irina CHUGUR et Juan M. CIGARRÁN : Indexing with wordnet synsets can improve text retrieval. *CoRR*, cmp-lg/9808002, 1998.

Brendan S. GUILLON : Ambiguity, generality, and indeterminacy : Tests and definitions. *Synthese*, 85(3):391–416, 1990.

Brendan S. GUILLON : *Ambiguity, indeterminacy, deixis and vagueness.*, pages 157–187. Oxford University Press, 2004.

Benoît HABERT, Gabriel ILLOUZ et Helka FOLCH : *Des décalages de distribution aux divergences d'acception*, pages 277–318. Lavoisier, 2005.

Marti HEARST : *User Interfaces for search*. Addison Wesley Professional, 2011.

Marti A. HEARST : Clustering versus faceted categories for information exploration. *Communications of the ACM*, 49:59–61, April 2006. ISSN 0001-0782.

Marti A. HEARST : *Search User Interfaces*. Cambridge University Press, 2009.

Jeff HUANG et Efthimis N. EFTHIMIADIS : Analyzing and evaluating query reformulation strategies in web search logs. *In Proceedings of the 18th ACM conference on Information and knowledge management*, CIKM '09, pages 77–86. ACM, 2009.

Maureen E. HUPFER et Brian DETLOR : Gender and web information seeking : a self-concept orientation model. *Journal of the American Society for Information Science and Technology*, 57(8):1105–1115, 2006.

Nancy IDE et Jean VÉRONIS : Word sense disambiguation : The state of the art. *Computational Linguistics*, 24:1–40, 1998.

Peter INGWERSEN : *Information Retrieval Interaction*. Taylor Graham, 1992.

Peter INGWERSEN : Polyrepresentation of information needs and semantic entities : Elements of a cognitive theory for information retrieval interaction. *In* W. Bruce CROFT et C. J. van RIJSBERGEN, éditeurs : *Proceedings of the 17th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, pages 101–110. ACM/Springer, 1994.

Peter INGWERSEN et Kalervo JÄRVELIN : Information retrieval in context : Sigir 2004 irix. pages 6–9. Sheffield University, 2004.

Peter INGWERSEN et Kalervo JÄRVELIN : *The Turn : Integration of Information Seeking and Retrieval in Context (The Information Retrieval Series)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2005. ISBN 140203850X.

Guillaume JACQUET et Fabienne VENANT : Construction automatique de classes de sélection distributionnelle. *In TALN 2005*, Dourdan, 2005.

Guillaume JACQUET, Fabienne VENANT et Bernard VICTORRI : *Polysémie lexicale*, pages 99–132. 2005.

Bernard J. JANSEN, Danielle L. BOOTH et Amanda SPINK : Determining the user intent of web search engine queries. *In Proceedings of the 16th international conference on World Wide Web*, WWW '07, pages 1149–1150, New York, NY, USA, 2007. ACM. ISBN 978-1-59593-654-7.

Bernard J. JANSEN, Amanda SPINK et Tefko SARACEVIC : Real life, real users, and real needs : A study and analysis of user queries on the web. *Information Processing and Management*, 36(2):207–227, 2000.

Kerstin JONASSON : *Le nom propre*. Champs linguistiques. De Boeck Supérieur, 1994.

Mika KÄKI : Optimizing the number of search result categories. *In CHI '05 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '05, pages 1517–1520. ACM, 2005.

Adam KILGARRIFF : *Word Senses*, pages 29–45. Springer, 2006.

Adam KILGARRIFF et Colin YALLOP : What's in a thesaurus. *In Proceedings of the Second Conference on Language Resources and Evaluation (LREC)*, pages 1371–1379, 2000.

Georges KLEIBER : *Problèmes de sémantique : la polysémie en questions*. Presses Universitaires du Septentrion, 1999.

Arnaud KLEIN : Wikipédia et la légitimité de la construction collective du savoir sur internet. 2005. URL http://www.internetactu.net/2005/05/25/wikipdia-et-la-lgitimit-de-la-construction-collective-du-savoir-sur-internet/.

Anders KOFOD-PETERSEN et Agnar AAMODT : Case-based situation assessment in a mobile context-aware system. *In Artificial Intelligence in Mobile Systems (AIMS 2003)*, pages 41–49, 2003.

Alice KRIEG-PLANQUE : A propos des "noms propres d'événement". Evénementialité et discursivité. *Les Carnets du Cediscor 11 : Le nom propre en discours.*, 2009.

Saul A. KRIPKE : *Naming and Necessity*. Harvard University Press, 1980.

Robert KROVETZ et W. Bruce CROFT : Lexical ambiguity and information retrieval. *ACM Transactions on Information Systems (TOIS)*, 10:115–141, 1992.

Giridhar KUMARAN et James ALLAN : Adapting information retrieval systems to user queries. *Information Processing Management*, 44(6):1838–1862, 2008.

Michelle LECOLLE : Toponymes en jeu : Diversité et mixage des emplois métonymiques de toponymes. *Studii si cercetari filologice*, 4:5–13, 2004.

Pierre LEFÈVRE : *La recherche d'informations*. Paris, 2000.

Michael LESK : Automatic sense disambiguation using machine readable dictionaries : how to tell a pine cone from an ice cream cone. *In Proceedings of the 5th annual international conference on Systems documentation*, SIGDOC '86, pages 24–26, New York, NY, USA, 1986. ACM.

Dekang LIN : Using syntactic dependency as local context to resolve word sense ambiguity. *In Proceedings of the eighth conference on European chapter of*

the Association for Computational Linguistics, EACL '97, pages 64–71, Stroudsburg, PA, USA, 1997. Association for Computational Linguistics.

Yoelle S. MAAREK, Ronald FAGIN, Israel Z. BEN-SHAUL et Dan PELLEG : Ephemeral document clustering for web applications. Rapport technique, IBM RESEARCH REPORT RJ 10186, 2000.

Christopher D MANNING, Prabhakar RAGHAVAN et Hinrich SCHÜTZE : *Introduction to information retrieval*, volume 1. Cambridge University Press Cambridge, 2008.

Gary MARCHIONINI et Ryen W. WHITE : Find what you need, understand what you find. *Journal of Human-Computer Interaction*, 23(3):205–237, 2008.

Oliver A. MCBRYAN : Genvl and wwww : Tools for taming the web. *In First International Conference on the World Wide Web. CERN, Geneva (Switzerland).*, 1994.

Olena MEDELYAN, Catherine LEGG, David N. MILNE et Ian H. WITTEN : Mining meaning from wikipedia. *International Journal of Human-Computer Studies*, 67(9):716–754, 2009.

Rada MIHALCEA : Using wikipedia for automatic word sense disambiguation. *In North American Chapter of the Association for Computational Linguistics (NAACL 2007)*, 2007.

Rada MIHALCEA, Paul TARAU et Elizabeth FIGA : Pagerank on semantic networks, with application to word sense disambiguation. *In Proceedings of the 20th international conference on Computational Linguistics*, COLING '04, Stroudsburg, PA, USA, 2004. Association for Computational Linguistics.

Rada F. MIHALCEA : Bootstrapping large sense tagged corpora. *In Proceedings of the 3rd International Conference on Language Resources and Evaluations (LREC), Las Palmas*, 2002.

George A. MILLER, Claudia LEACOCK, Randee TENGI et Ross T. BUNKER : A semantic concordance. *In Proceedings of the workshop on Human Language Technology*, HLT '93, pages 303–308. Association for Computational Linguistics, 1993.

Stefano MIZZARO : Relevance : The whole (hi)story. *Journal of the American Society for Information Science*, 48(9):810–832, 1997.

Stefano MIZZARO : How many relevances in information retrieval ? *Interacting With Computers*, 10:305–322, 1998.

Jacques MOESCHLER et Anne REBOUL : *Dictionnaire encyclopédique de pragmatique*. 1994.

Véronique MORICEAU et Patrick SAINT-DIZIER : Métaphore. *In* Danièle GODARD et Francis CORBLIN, éditeurs : *Sémanticlopédie : dictionnaire de sémantique*. GDR Sémantique et Modélisation, CNRS, http ://www.semantique-gdr.net/dico/, 2006.

Peter MORVILLE et Louis ROSENFELD : *Information architecture for the World Wide Web*. 1998.

Josiane MOTHE : *Recherche d'information contextuelle : le cas des requêtes*. Lavoisier, 2011.

Josiane MOTHE et Ludovic TANGUY : Linguistic features to predict query difficulty. *In ACM SIGIR 2005 Workshop on Predicting Query Difficulty - Methods and Applications*, 2005.

Jori MUR : Increasing the coverage of answer extraction by applying anaphora resolution. *In Fifth Slovenian and First International Language Technologies Conference (IS-LTC '06)*, 2006.

Emmanuel NAVARRO, Yannick CHUDY, Bruno GAUME, Guillaume CABANAC et Karen PINEL-SAUVAGNAT : Kodex ou comment organiser les résultats d'une recherche d'information par détection de communautés sur un graphe biparti ? *In CORIA'11, Avignon*, pages 25–40. ARIA, mars 2011.

Raquel NAVARRO-PRIETO, Mike SCAIFE et Yvonne ROGERS : Cognitive strategies in web searching. *In Proceedings of the 5th Conference on Human Factors the Web*, pages 1–13, 1999.

Roberto NAVIGLI : Word sense disambiguation : A survey. *ACM Computing Surveys (CSUR)*, 41(2):10–69, 2009.

Adeline NAZARENKO : *Sur quelle sémantique repose les méthodes automatiques d'accès au contenu textuel ?*, pages 211–244. Lavoisier, 2005.

Hwee Tou NG et Hian Beng LEE : Integrating multiple knowledge sources to disambiguate word sense : an exemplar-based approach. *In Proceedings of the 34th annual meeting on Association for Computational Linguistics*, ACL '96, pages 40–47, Stroudsburg, PA, USA, 1996. Association for Computational Linguistics.

David NICOLAS : Ambiguïté. *In* Danièle GODARD, Laurent ROUSSARIE et Francis CORBLIN, éditeurs : *Sémanticlopédie : dictionnaire de sémantique*. GDR Sémantique et Modélisation, CNRS, http ://www.semantique-gdr.net/dico/, 2006. URL http://www.semantique-gdr.net/dico/index.php/Ambigu%C3%AFt%C3%A9.

Jacob NIELSEN : Usability 101. 2003. URL http://www.useit.com/alertbox/20030825.html.

Geoffrey NUNBERG : Transfers of meaning. *Journal of Semantics*, 12:109–132, 1995.

Stanislaw OSINSKI, Jerzy STEFANOWSKI et Dawid WEISS : Lingo : Search results clustering algorithm based on singular value decomposition. *In Proceedings of the International Intelligent Information Processing and Web Mining Conference.*, Advances in Soft Computing., pages 359–368. Springer, 2004.

Iadh OUNIS, Christina LIOMA, Craig MACDONALD et Vassilis PLACHOURAS : Research directions in terrier. *Novatica/UPGRADE Special Issue on Web Information Access, Ricardo Baeza-Yates et al. (Eds), Invited Paper*, 2007.

Patrick PANTEL et Dekang LIN : Discovering word senses from text. *In Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '02, pages 613–619, New York, NY, USA, 2002. ACM.

Marie-Anne PAVEAU : Le toponyme, désignateur souple et organisateur mémoriel. L'exemple du nom de bataille. *Mots. Les langages du politique*, 86:23–35, 2008.

Ted PEDERSEN et Rebecca BRUCE : Distinguishing word senses in untagged text. *In Proceedings of the Second Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 197–207, 1997.

Aurélie PICTON : *Diachronie en langue de spécialité. Définition d'une méthode linguistique outillée pour repérer l'évolution des connaissances en corpus. Un exemple appliqué au domaine spatial. Thèse de doctorat en Sciences du Langage*. Thèse de doctorat, Université Toulouse 2, 2009.

Karen PINEL-SAUVAGNAT : *Modèle flexible pour la recherche d'information dans des corpus de documents semi-structurés*. Thèse de doctorat, Université Paul Sabatier, juin 2005.

James PUSTEJOVSKY : *The generative lexicon*. The MIT Press, 1995.

R. RADA, H. MILI, E. BICKNELL et M. BLETTNER : Development and application of a metric on semantic nets. *IEEE Trans. Syst. Man Cybernet*, 19(1):17–30, 1989.

Mandar A. RAHURKAR, Dan ROTH et Thomas S. HUANG : Which "apple" are you talking about ? *In Proceeding of the 17th international conference on World Wide Web*, WWW '08, pages 1197–1198, New York, NY, USA, 2008. ACM. ISBN 978-1-60558-085-2.

William M. RAND : Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, 66(336):846–850, 1971.

Franck REBILLARD : Du traitement de l'information à son retraitement. La publication de l'information journalistique sur l'internet. *Réseaux*, 137(3): 29–68, 2006.

Franck REBILLARD et Nikos SMYRNAIOS : Les infomédiaires, au coeur de la filière de l'information en ligne. Le cas de google, wikio et paperblog. *Réseaux*, 160-161:163–194, 2010.

Martin RIEGEL, Jean-Christophe PELLAT et René RIOUL : *Grammaire Méthodique du Français*. Presses Universitaires de France, 1998.

Soo Young RIEH et Hong Iris XIE : Analysis of multiple query reformulations on the web : The interactive information retrieval context. *Information Processing Management*, 42(3):751–768, 2006.

Stephen E ROBERTSON : The probability ranking principle in IR. *Journal of documentation*, 33(4):294–304, 1977.

Stephen E. ROBERTSON et Stephen WALKER : Some simple effective approximations to the 2-Poisson model for probabilistic weighted retrieval. *In SIGIR '94 : Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 232–241. Springer-Verlag New York, Inc., 1994.

Peter Mark ROGET : *Roget's International Thesaurus.* 1st ed. Cromwell., 1911.

Gerard SALTON : A comparison between manual and automatic indexing methods. Rapport technique, 1968.

Gerard SALTON, Edward A. FOX et Harry WU : Extended boolean information retrieval. *Commun. ACM*, 26(11):1022–1036, 1983.

Gerard SALTON et Michael J. MCGILL : *Introduction to Modern Information Retrieval*. McGraw-Hill, Inc., New York, NY, USA, 1986. ISBN 0070544840.

Mark SANDERSON : Word sense disambiguation and information retrieval. *In Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '94, pages 142–151, New York, NY, USA, 1994. Springer-Verlag New York, Inc.

Mark SANDERSON : *Word Sense Disambiguation and information retrieval*. Thèse de doctorat, Department of Computing Science at University of Glasgow, Glasgow G12 8QQ, UK, 1997.

Mark SANDERSON : Retrieving with good sense. *Information Retrieval*, 2(1):45–65, 2000.

Mark SANDERSON : Ambiguous queries : test collections need more sense. *In SIGIR*, pages 499–506, 2008.

Mark SANDERSON et Susan DUMAIS : Examining repetition in user search behavior. *In Proceedings of the 29th European conference on IR research*, ECIR'07, pages 597–604. Springer-Verlag, 2007.

Celina SANTAMARÍA, Julio GONZALO et Javier ARTILES : Wikipedia as sense inventory to improve diversity in web search results. *In ACL '10 Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1357–1366, 2010.

Rodrygo L. T. SANTOS, Jie PENG, Craig MACDONALD et Iadh OUNIS : Explicit search result diversification through sub-queries. *In Proceedings of the 32nd European conference on Advances in Information Retrieval*, ECIR'2010, pages 87–99. Springer-Verlag, 2010a.

Rodrygo L.T. SANTOS, Craig MACDONALD et Iadh OUNIS : Exploiting query reformulations for web search result diversification. *In Proceedings of the 19th international conference on World wide web*, WWW '10, pages 881–890. ACM, 2010b.

Tefko SARACEVIC : Effects of inconsistent relevance judgments on information retrieval test results : A historical perspective. *Library Trends*, 56(4):763–783, 2008.

Hinrich SCHÜTZE : Dimensions of meaning. *In Proceedings of the 1992 ACM/IEEE conference on Supercomputing*, Supercomputing '92, pages 787–796, Los Alamitos, CA, USA, 1992. IEEE Computer Society Press.

Hinrich SCHÜTZE et Jan O. PEDERSEN : Information retrieval based on word senses. *In Symposium on Document Analysis and Information Retrieval*, 1995.

Xuehua SHEN et ChengXiang ZHAI : Exploiting query history for document ranking in interactive information retrieval. *In Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, SIGIR '03, pages 377–378. ACM, 2003.

Ben SHNEIDERMAN, Don BYRD et W. B CROFT : Clarifying search : A user-interface framework for text searches. Rapport technique, 1997.

Ben SHNEIDERMAN et Catherine PLAISANT : *Designing the user interface : strategies for effective human-computer interaction, 4ème édition*. Addison Wesley, 2004.

Paul SIBLOT : De la signifiance du nom propre. *Cahiers de praxématique*, 8:97–114, 1987.

Ruihua SONG, Zhenxiao LUO, Jian-Yun NIE, Yong YU et Hsiao-Wuen HON : Identification of ambiguous queries in web search. *Information Processing and Management*, 45(2):216–229, 2009.

Karen SPÄRCK-JONES, Stephen E. ROBERTSON et Mark SANDERSON : Ambiguous requests : implications for retrieval tests, systems and theories. *SIGIR Forum*, 41(2):8–17, décembre 2007.

Amanda SPINK, Bernard J. JANSEN et Jan PEDERSEN : Searching for people on web search engine. *Journal of Documentation*, 60(3):266–278, 2004.

Amanda SPINK, Bernard J. JANSEN, Dietmar WOLFRAM et Tefko SARACEVIC : Characteristics of question format web queries : an exploratory study. *Information Processing Management*, 38(4):453–471, 2002a.

Amanda SPINK, Bernard J. JANSEN, Dietmar WOLFRAM et Tefko SARACEVIC : From e-sex to e-commerce : Web search changes. *Computer*, 35(3):107–109, 2002b.

Christopher STOKOE : Automated word sense disambiguation for web information retrieval. *SIGIR Forum*, 39(1):68–68, juin 2005.

Christopher STOKOE, Michael P. OAKES et John TAIT : Word sense disambiguation in information retrieval revisited. *In SIGIR*, pages 159–166, 2003.

Alistair SUTCLIFFE et Mark ENNIS : Towards a cognitive theory of information retrieval. *Interacting with Computers*, 10(3):321–351, 1998.

Russell SWAN et David JENSEN : Timemines : Constructing timelines with statistical models of word usage. *In Proceedings of the 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD 2000, pages 73–80, 2000.

Lynda TAMINE-LECHANI : *De la recherche d'information orientée système vers la recherche d'information orientée contexte : Verrous, contributions, perspectives*. Habilitation à diriger des recherches, Université Paul Sabatier, 2008.

Robert S. TAYLOR : Question-negotiation and information seeking in librairies. 29:178–194, 1968.

Jaime TEEVAN, Eytan ADAR, Rosie JONES et Michael POTTS : History repeats itself : repeat queries in yahoo's logs. *In Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '06, pages 703–704. ACM, 2006.

Jaime TEEVAN, Susan T. DUMAIS et Eric HORVITZ : Beyond the commons : Investigating the value of personalizing web search. *In Proceedings of the Workshop on New Technologies for Personalized Information Access (PIA)*, pages 84–92, 2005.

Jaime TEEVAN, Susan T. DUMAIS et Daniel J. LIEBLING : To personalize or not to personalize : modeling queries with variation in user intent. *In Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '08, pages 163–170, New York, NY, USA, 2008. ACM.

Anastasios TOMBROS, Ian RUTHVEN et Joemon M. JOSE : How users assess web pages for information seeking. *Journal of American Society of Information Science and Technology (JASIST)*, 56(4):327–344, 2005.

Peter TURNEY : Word sense disambiguation by web mining for word co-occurrence probabilities. *In Proceedings of the 3eme international conference "Evaluation of Systems for the Semantic Analysis of Text"*, SENSEVAL-3 2004, 2004.

Sarah K. TYLER et Jaime TEEVAN : Large scale query log analysis of re-finding. *In Proceedings of the third ACM international conference on Web search and data mining*, WSDM '10, pages 191–200. ACM, 2010.

Lonneke van der PLAS : *Automatic lexico-semantic acquisition for question answering*. Thèse de doctorat, GRODIL, 2008.

Bernard VICTORRI et Catherine FUCHS : *La polysémie. Construction dynamique du sens*. Hermès, 1996.

Ellen M. VOORHEES : Using wordnet to disambiguate word senses for text retrieval. *In SIGIR*, pages 171–180, 1993.

Jean VÉRONIS : Sense tagging : does it make sense. *In Corpus linguistics'*, 2001.

Jean VÉRONIS : Les dictionnaires traditionnels sont-ils adaptés au traitement du sens en T.A.L. ? . *In Journée d'étude de l'ATALA "Les dictionnaires électroniques"*, Paris, 2002.

Jean VÉRONIS : Cartographie lexicale pour la recherche d'information. *In Actes de la 10ème Conférence sur le Traitement Automatique des Langues Naturelles (TALN), 11-14 juin 2003*, pages 265–274, 2003.

Steve WALKER, Stephen E. ROBERTSON, Mohand BOUGHANEM, Gareth J. F. JONES et Karen SPARCK JONES : Okapi at TREC-6 Automatic ad hoc, VLC, routing, filtering and QSDR. *NIST SPECIAL PUBLICATION SP*, pages 125–136, 1998.

Yu WANG et Eugene AGICHTEIN : Query ambiguity revisited : clickthrough measures for distinguishing informational and ambiguous queries. *In Human Language Technologies : The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, pages 361–364. Association for Computational Linguistics, 2010.

Stephen F. WEISS : Learning to disambiguate. *Information Storage and Retrieval*, 9(1):33–41, 1973.

Michael J WELCH, Junghoo CHO et Christopher OLSTON : Search result diversity for informational queries. *In Proceedings of the 20th international conference on World wide web (WWW)*, pages 237–246. ACM, 2011.

Gui-Rong XUE, Dikan XING, Qiang YANG et Yong YU : Deep classification in large-scale text hierarchies. *In Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '08, pages 619–626, New York, NY, USA, 2008. ACM.

David YAROWSKY : Word-sense disambiguation using statistical models of roget's categories trained on large corpora. *In Proceedings of the 14th conference on Computational linguistics*, volume 2 de *COLING '92*, pages 454–460, Stroudsburg, PA, USA, 1992. Association for Computational Linguistics.

David YAROWSKY : One sense per collocation. *In Proceedings of the workshop on Human Language Technology*, HLT '93, pages 266–271, Stroudsburg, PA, USA, 1993. Association for Computational Linguistics.

David YAROWSKY : Unsupervised word sense disambiguation rivaling supervised methods. *In Proceedings of the 33rd annual meeting on Association for*

Computational Linguistics, ACL '95, pages 189–196, Stroudsburg, PA, USA, 1995. Association for Computational Linguistics.

Oren ZAMIR et Oren ETZIONI : Web document clustering : a feasibility demonstration. *In Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '98, pages 46–54, New York, NY, USA, 1998. ACM.

Oren ZAMIR, Oren ETZIONI, Omid MADANI et Richard M. KARP : Fast and intuitive clustering of web documents. *In Proceedings of the 3rd International Conference on Knowledge Discovery and Data Mining*, pages 287–290, 1997.

Hua-Jun ZENG, Qi-Cai HE, Zheng CHEN, Wei-Ying MA et Jinwen MA : Learning to cluster web search results. *In Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '04, pages 210–217. ACM, 2004.

Cheng Xiang ZHAI, William W. COHEN et John LAFFERTY : Beyond independent relevance : methods and evaluation metrics for subtopic retrieval. *In Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, SIGIR '03, pages 10–17, 2003.

ChengXiang ZHAI et John LAFFERTY : A risk minimization framework for information retrieval. *Information Processing and Management : an International Journal - Special issue : Formal methods for information retrieval*, 42(1):31–55, 2006.

Benyu ZHANG, Hua LI, Yi LIU, Lei JI, Wensi XI, Weiguo FAN, Zheng CHEN et Wei-Ying MA : Improving web search results using affinity graph. *In Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '05, pages 504–511, New York, NY, USA, 2005. ACM.

Dell ZHANG et Jinsong LU : What queries are likely to recur in web search ? *In Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '09, pages 827–828. ACM, 2009.

Xiaodan ZHANG, Xiaohua HU et Xiaohua ZHOU : A comparative evaluation of different link types on enhancing document clustering. *In Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '08, pages 555–562, New York, NY, USA, 2008. ACM.

Annexes

Annexe A

Données

A.1 Liste des urls des moteurs spécialisés

Ces moteurs sont présentés dans le chapitre 4.
- Cluzz : http://www.cluuz.com/
- Tile.net : http://tile.net/
- 1001Forums : http://www.1001forums.fr/
- Wikio : http://www.wikio.fr/
- Twitter : https://twitter.com/search
- Topsy :http://topsy.com/
- Pixolution : http://www.pixolution.de
- Dailymotion : http://www.dailymotion.com/
- PdfGeni : http://www.pdfgeni.com/index.php
- Scirius : http://www.scirus.com/
- CiteseerX : http://citeseerx.ist.psu.edu/
- EMM : http://emm.newsbrief.eu/NewsBrief/clusteredition/fr/latest.html
- GoogleNews : http://news.google.fr/

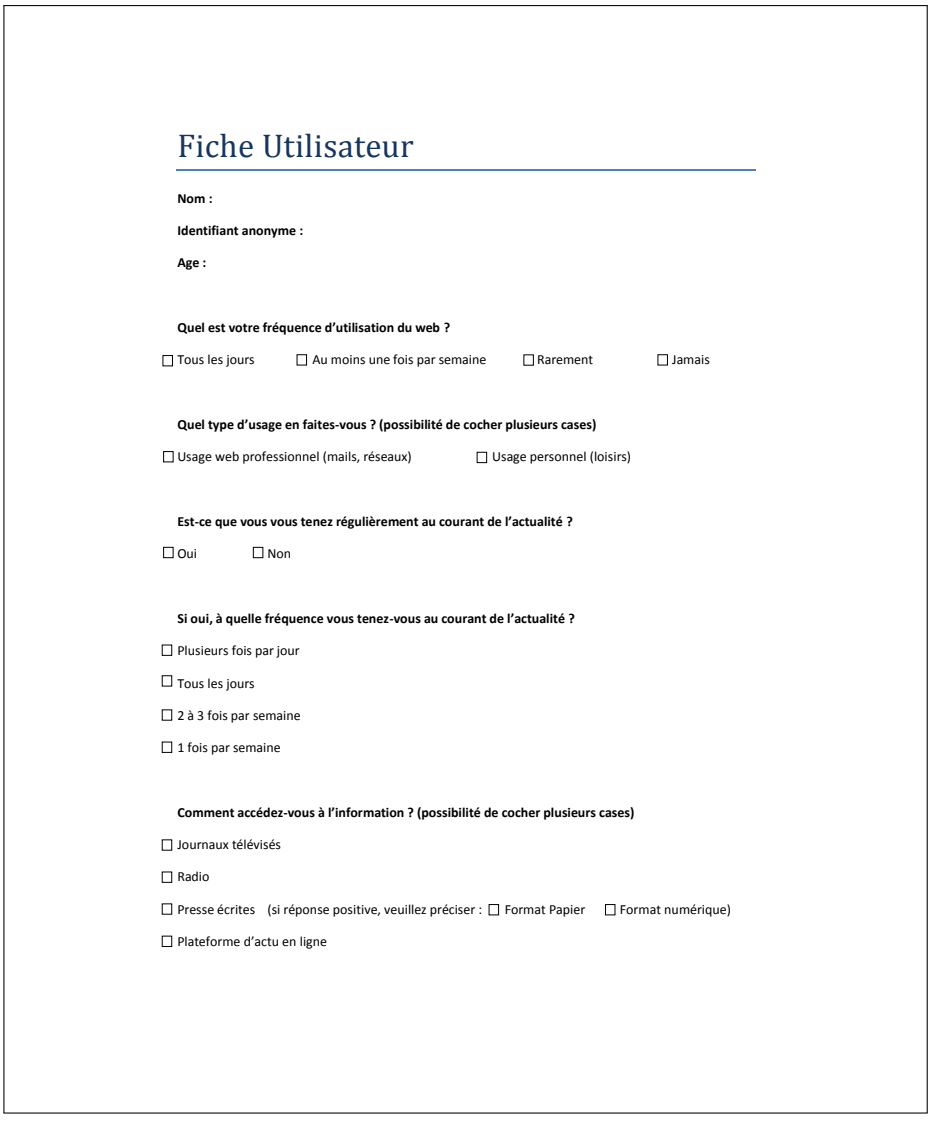
A.2 Corpus 2424reqFréquentes

Le corpus 2424reqFréquentes n'est pas disponible.

Annexe B

Documents complémentaires pour les tests utilisateurs

B.1 Questionnaire expérience 1



Fiche Utilisateur

Nom :

Identifiant anonyme :

Age :

Quel est votre fréquence d'utilisation du web ?

☐ Tous les jours ☐ Au moins une fois par semaine ☐ Rarement ☐ Jamais

Quel type d'usage en faites-vous ? (possibilité de cocher plusieurs cases)

☐ Usage web professionnel (mails, réseaux) ☐ Usage personnel (loisirs)

Est-ce que vous vous tenez régulièrement au courant de l'actualité ?

☐ Oui ☐ Non

Si oui, à quelle fréquence vous tenez-vous au courant de l'actualité ?

☐ Plusieurs fois par jour

☐ Tous les jours

☐ 2 à 3 fois par semaine

☐ 1 fois par semaine

Comment accédez-vous à l'information ? (possibilité de cocher plusieurs cases)

☐ Journaux télévisés

☐ Radio

☐ Presse écrites (si réponse positive, veuillez préciser : ☐ Format Papier ☐ Format numérique)

☐ Plateforme d'actu en ligne

FIGURE B.1 : Questionnaire rempli par les sujets lors de la passation de leur test

B.2 Résultats des tests utilisateurs

B.2.1 Expérience 1

Les tableaux B.2 à B.10 retranscrivent les résultats en terme de réponses vides et non vides à l'expérience 1 (cf. 7.1). Chaque tableau résume les résultats pour chaque requête du panel test. Les réponses dites « exploitables » correspondent à la soustraction des réponses identiques à la requête aux réponses non vides.

afghanistan	Réponses exploitables
INTERNATIONAL	20 / 20
SOCIÉTÉ	17 / 20
CULTURES	20 / 20
total	57 / 60

TABLEAU B.2 : Résultats requête *afghanistan*

wikileaks	Réponses exploitables
INTERNATIONAL	16 / 20
SOCIÉTÉ	20 / 20
CULTURES	18 / 20
ECONOMIE	20 / 20
total	74 / 80

TABLEAU B.3 : Résultats requête *wikileaks*

berlusconi	Réponses exploitables
INTERNATIONAL	19 / 20
CULTURES	20 / 20
ECONOMIE	9 / 20
total	48 / 60

TABLEAU B.4 : Résultats requête *berlusconi*

tunisie	Réponses exploitables
INTERNATIONAL	16 / 20
SOCIÉTÉ	12 / 20
CULTURES	9 / 20
ECONOMIE	20 / 20
POLITIQUE	19 / 20
total	76 / 100

TABLEAU B.5 : Résultats requête *tunisie*

leatitia	Réponses exploitables
POLITIQUE	18 / 20
SOCIÉTÉ	20 / 20
CULTURES	4 / 20
ECONOMIE	0 / 20
total	42 / 80

TABLEAU B.6 : Résultats requête *leatitia*

grève	Réponses exploitables
INTERNATIONAL	10 / 20
SOCIÉTÉ	17 / 20
CULTURES	6 / 20
ECONOMIE	17 / 20
POLITIQUE	13 / 20
SPORT	20 / 20
total	83 / 120

TABLEAU B.7 : Résultats requête *grève*

égypte	Réponses exploitables
INTERNATIONAL	19 / 20
SOCIÉTÉ	17 / 20
POLITIQUE	16 / 20
ECONOMIE	9 / 20
total	61 / 80

TABLEAU B.8 : Résultats requête *egypte*

médicaments	Réponses exploitables
INTERNATIONAL	3 / 20
SOCIÉTÉ	13 / 20
CULTURES	17 / 20
ECONOMIE	16 / 20
total	49 / 80

TABLEAU B.9 : Résultats requête *médicaments*

météo	Réponses exploitables
INTERNATIONAL	20 / 20
SOCIÉTÉ	20 / 20
CULTURES	6 / 20
total	46 / 20

TABLEAU B.10 : Résultats requête *météo*

B.2.2 Expérience 2

Le tableau B.11 regroupe les résultats de l'évaluation de l'expérience 2 en 7.2.4.

	user2		user4		user5		user6		user7		Moyenne	
Requêtes	ARI	RI	ARI	RI	ARI	RI	ARI	RI	ARI	RI	ARI	RI
afghanistan	0,42	0,73	0,42	0,73	0,24	0,64	0,22	0,65	0,51	0,77	0,36	0,70
berlusconi	0,00	0,58	0,07	0,45	0,00	0,58	0,01	0,52	0,06	0,56	0,02	0,53
égypte	0,01	0,53	0,01	0,61	0,04	0,64	0,10	0,70	0,07	0,67	0,04	0,63
grève	0,36	/	0,31	0,82	0,23	0,75	0,31	0,82	0,31	0,82	0,30	/
laetitia	0,33	0,73	0,37	0,78	0,13	0,54	0,44	0,81	0,21	0,67	0,29	0,705
médicaments	0,17	0,65	0,22	0,71	0,12	0,64	0,07	0,50	0,10	0,69	0,13	0,63
météo	0,86	0,86	0,53	0,86	0,86	0,86	0,71	0,82	0,78	0,85	0,74	0,85
tunisie	0,16	0,75	0,24	0,76	0,16	0,70	0,20	0,74	0,35	0,81	0,22	0,75
wikileaks	0,46	0,82	0,46	0,82	0,48	0,83	0,28	0,67	0,59	0,82	0,45	0,78

TABLEAU B.11 : Résultats des Adjusted Rank Index et Rand Index

Etude de l'ambiguïté des requêtes dans un moteur de recherche spécialisé dans l'actualité : exploitation d'indices contextuels

Résumé

Dans cette thèse, nous envisageons la question de l'ambiguïté des requêtes soumises à un moteur de recherche dans un domaine particulier qui est l'actualité. Nous nous appuyons sur les travaux récents dans le domaine de la recherche d'information (RI) qui ont montré l'apport d'informations contextuelles pour mieux cerner et traiter plus adéquatement le besoin informationnel. Nous faisons ainsi l'hypothèse que les éléments d'information disponibles dans une application de RI (contextes présents dans la base documentaire, répétitions et reformulations de requêtes, dimension diachronique de la recherche) peuvent nous aider à étudier ce problème d'ambiguïté. Nous faisons également l'hypothèse que l'ambiguïté va se manifester dans les résultats ramenés par un moteur de recherche. Dans ce but, nous avons mis en place un dispositif pour étudier l'ambiguïté des requêtes reposant sur une méthode de catégorisation thématique des requêtes, qui s'appuie sur une catégorisation experte. Nous avons ensuite montré que cette ambiguïté est différente de celle repérée par une ressource encyclopédique telle que Wikipédia. Nous avons évalué ce dispositif de catégorisation en mettant en place deux tests utilisateurs. Enfin, nous fournissons une étude basée sur un faisceau d'indices contextuels afin de saisir le comportement global d'une requête.

Mots-clés : Recherche d'information, Analyse des requêtes, Actualités, Ambiguïté, Traitement Automatique des Langues

Study of the ambiguity of queries in a news search engine : exploitation of contextual clues

Abstract

In this thesis, we consider the question of the ambiguity of queries submitted to a search engine in a particular area that is news. We build on recent work in the field of information retrieval (IR) that showed the addition of contextual information to better identify and address more adequately the information need. On this basis, we make the hypothesis that the elements of information available in an application of IR (contexts in the collection of documents, repetitions and reformulations of queries, diachronic dimension of the search) can help us to examine this problem of ambiguity. We also postulate that ambiguity will manifest in the results returned by a search engine. In this purpose to evaluate these hypotheses, we set up a device to study the ambiguity of queries based on a method of thematic categorization of queries, which relies on an expert categorization. We then show that this ambiguity is different which is indicated by an encyclopedic resources such as Wikipedia. We evaluate this categorization device by setting up two user tests. Finally, we carry out a study based on a set of contextual clues in order to understand the global behavior of a query.

Mots-clés en anglais : Information Retrieval, Analysis of queries, News, Ambiguity, Natural language processing